Nelson Alves Morais

Clustering de relacionamentos entre entidades nomeadas em textos com base no contexto

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
4 de Novembro de 2016

Nelson Alves Morais

Clustering de relacionamentos entre entidades nomeadas em textos com base no contexto

Dissertação submetida à Faculdade de Ciências da Universidade do Porto como parte dos requisitos para a obtenção do grau de Mestre em Ciência de Computadores

Orientador: Professor Doutor Alípio Mário Guedes Jorge
Coorientadora: Doutora Maria da Conceição de Oliveira Nunes Rocha

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
4 de Novembro de 2016

Agradecimentos

Ao orientador, Prof. Doutor Alípio Jorge, por toda a disponibilidade em ajudar a melhorar o meu trabalho. O seu contributo foi fundamental e imprescindível.

À coorientadora, Doutora Conceição Rocha, por acrescentar outra perspetiva e ter colaborado em etapas cruciais do trabalho.

Aos Sapo Labs, por terem disponibilizado os textos.

Aos meus pais, principalmente por toda a assistência prestada.

A outros familiares e também amigos.

A todos os que contribuiram para a minha inclusão durante todo o meu percurso escolar.

Finalmente, a todos os professores com quem tive o privilégio de aprender.

Projeto "TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020" foi financiado pelo Programa Operacional Regional do Norte (NORTE 2020), sobre o Acordo de Parceria PORTUGAL 2020, e através do Fundo Europeu de Desenvolvimento Regional (FEDER).

Resumo

Um problema importante na compreensão automática de texto é a identificação e caracterização de relações entre entidades nomeadas em textos. Nesta dissertação, é apresentada uma abordagem baseada no contexto que caracteriza relações entre entidades nos textos. As entidades são identificadas primeiramente por um algoritmo de Reconhecimento de Entidades Nomeadas (REN) - neste caso é usado Pampo[1]. Num segundo passo, são identificadas as frases onde duas entidades ocorrem exatamente e é extraído o contexto (palavras envolventes) de cada par de entidades. As palavras no contexto ajudarão a fornecer semântica para cada uma das relações identificadas. Finalmente, é utilizada uma técnica de agrupamento para identificar pares de entidades semanticamente semelhantes e caracterizar esses relacionamentos usando palavras selecionadas no contexto.

Palavras-chave: *estatística*, *inteligência artificial*, *dados não estruturados*, *dados semi-estruturados*, *mineração de texto*, *entidades nomeadas*, *relações semânticas*, *processamento de contexto*, *medida de similaridade*, *clustering*, *etiquetagem*, *estrutura de dados*

Abstract

One important problem in automatic text understanding is the identification and characterization of relations between entities named in the text. In this dissertation, it is presented a context-based approach which characterizes relations between entities from texts. Entities are first identified by a Named Entity Recognition (NER) algorithm - in this case it is used Pampo[1]. In a second step, it are identified the sentences where exactly two entities occur and extract the context (surrounding words) of each pair of entities. The words in the context will help provide semantics to each of the identified relations. Finally, it is employed a clustering technique to identify semantically similar pairs of entities and characterize those relationships using selected words in the context.

Keywords: *statistics*, *artificial intelligence*, *unstructured data*, *semi-structured data*, *text mining*, *named entities*, *semantic relations*, *context processing*, *similarity measure*, *clustering*, *labeling*, *data structure*

Conteúdo

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas

EN ..... Entidades nomeadas

END ..... Entidades nomeadas desinteressantes

PLN ..... Processamento de linguagem natural

RS ..... Relações semânticas

Capítulo 1

Introdução

A partir do século XX, a informatização tem sido uma constante e prevê-se que mantenha tal tendência[6]. Informação que antes era guardada em papel ou noutros suportes físicos, passou a estar no suporte digital. Fotografias, documentos, música, filmes e jogos são exemplos da mudança.

Por outro lado, a evolução tecnológica dos servidores e o acesso generalizado à Internet contribui de forma significativa para o aumento da distribuição e partilha de informação digital. Para que a informação possa ser facilmente consultada pelo utilizador, ela é guardada em bases de dados com uma estrutura pré-determinada. No entanto, nem sempre a forma dessa estrutura beneficia quer a transmissão quer a consulta de informação contida nos dados.

Com o objetivo de se tirar o máximo partido de dados disponíveis, que podem estar registados mas insuficientemente aproveitados, surgiu o *data mining* (mineração de dados). Este processo computacional envolve métodos na interceção de Estatística, Inteligência Artificial, Sistemas de Base de Dados, Aprendizagem Automática e Reconhecimento de Padrões. Há duas capacidades que podem ser aproveitadas neste processo: previsão e descrição. Tarefas preditivas são a classificação, a regressão e a deteção de anomalias. Enquanto tarefas descritivas são o *clustering*, a descoberta de regras de associação e a descoberta de padrões sequenciais.[7]

O crescimento exponencial da informação na Web é representado por notícias, relatórios, artigos científicos, postagens, etc. Este tipo de informação tem forma não

estruturada ou semi-estruturada, o que obriga o utilizador a uma interpretação da leitura dos textos para poder aceder à informação neles contida.

Por vezes, os textos encontram-se com anotações relativas a entidades nomeadas[8] (EN). Estas podem ser definidas como nomes que identificam tudo o que é concreto. E enquadram-se em diferentes domínios como pessoas, organizações, etc.

De forma a que essa informação seja aproveitada, a solução pode ser o processamento de documentos textuais, alocados na Intra/Internet. Para o desempenho de tal tarefa, existe o *web content mining*. Este processo é um cruzamento entre dois ramos (*web mining* e *text mining*), e envolve tarefas como classificação e *clustering* dos documentos.

Classificação entende-se como a atribuição de uma classe, entre um predefinido conjunto de classes, a cada instância com base nos seus atributos, dum determinado conjunto de dados. É entendida como *clustering* a divisão por grupos de instâncias, dum conjunto de dados, com base nos seus atributos.

Relações semânticas[9] (RS) são ligações entre significados. A aglomeração destas entre pares de EN, nos textos, é um exemplo duma tarefa de *clustering*. Uma vez concretizada, é permitido o conhecimento dos diferentes tipos de RS que se encontram descritas nos textos de entrada.

Especificamente, percebe-se a ocorrência de tipos em que a pessoa A trabalha para a organização B e a pessoa C trabalha para a organização D. Ou a existência da organização E ter participado no evento F.

1.1 Motivação

As RS entre pares de EN podem ser vistas como consultas a uma base de dados relacional[10]. A partir daqui, é possível agrupar pares com RS semelhantes se existirem dados que as descrevam. O conteúdo de quaisquer tabelas pode ser inserido diretamente por humanos. Também há a alternativa da inserção automática, embora possa haver pré-processamento do conteúdo inserido.

Se se pretender então que RS obtidas a partir dum conjunto de documentos textuais, com um conjunto de EN reconhecidas, sejam dispostas numa base de dados, é necessário que se processe a informação de modo a enquadrar-se nos esquemas das tabelas.

A resolução do problema da passagem de dados em textos para estruturas como tabelas pode passar pelo processamento de linguagem natural[11] (PLN) com uso de *part-of-speech tagging*[12] (POST). Depois da análise se concretizar, é importante saber o significado das palavras que pertencem a uma árvore sintática para que se confirme a existência de relacionamentos e quais.

A partir do momento em que duas ou mais EN se situam numa mesma frase ou num mesmo parágrafo, existe a hipótese de haver RS entre quaisquer pares destas. Por exemplo, a escolha, entre as palavras que estão no meio dos elementos dum par e todas as palavras da mesma frase, pode influenciar a qualidade dum resultado.

Há sítios com diferentes serviços que partilham tantos dados estruturados como dados semi-estruturados e/ou não estruturados. Exemplos disto são portais como o Zero Zero[1], o IMDB[2], a Wikipédia[3], o Público[4] e o Blogspot[5]. Portanto, encontram-se nestes casos a oportunidade de se aproveitarem os textos não estruturados ou semi-estruturados como complemento aos dados restantes ou como novas utilidades.

É importante, para a compreensão do mundo real, saber que pares podem ter RS semelhantes a partir de texto. De forma a que seja identificado cada grupo de pares, com tal semelhança, o uso de etiquetas é uma opção para a distinção. Todas as palavras que formam o contexto dos pares poderão influenciar a aglomeração consoante a semelhança que existe entre diferentes contextos, enquanto as palavras que caracterizam um relacionamento distinto poderão servir para etiquetagem.

Para as similaridades entre contextos serem medidas, é necessário ter em conta os termos que os constituem. O contexto pode ser visto como uma frase, um parágrafo

[1]http://www.zerozero.pt/ - Acedido em 03/11/2016
[2]http://www.imdb.com/ - Acedido em 03/11/2016
[3]https://en.wikipedia.org/ - Acedido em 03/11/2016
[4]http://www.publico.pt/ - Acedido em 03/11/2016
[5]https://www.blogger.com/ - Acedido em 03/11/2016

ou um conjunto de textos onde se inserem EN. Num caso específico, o contexto é o conjunto de todas as palavras, que não constituam EN, duma frase:

> *A **China** e a **Rússia** são dois dos países mais poderosos do planeta tal como os **Estados Unidos da América**.*

Neste caso, *China*, *Rússia* e *Estados Unidos da América* são termos que representam EN. Os restantes termos compõem o contexto.

Um exemplo prático onde a identificação e aglomeração de RS pode ser útil é nas redes complexas, que têm sido uma forma de mostrar distribuições de conexões existentes entre conceitos[13].

Redes sociais, entre as quais o Facebook[6], conseguem representar informação na forma destas estruturas através de ligações criadas entre as suas páginas pessoais e públicas. A descoberta de RS em dados textuais contribuirá para o crescimento das redes, reproduzindo as suas propriedades estruturais. Também esta descoberta de RS contribuirá beneficamente para outras aplicações como sistemas de recomendações de EN semelhantes, extração de conhecimento e recuperação de informação.

1.2 Objetivos

Existem dois principais objetivos a atingir neste trabalho de dissertação:

1. Desenvolver um algoritmo que consiga identificar e agrupar RS entre EN.
2. Testar o algoritmo e avaliar os resultados obtidos.

De modo a ser alcançado o 1º objetivo, há duas tarefas principais a executar. Em primeiro lugar, há que selecionar e integrar um método de reconhecimento de EN. A definição desta tarefa surge da necessidade de serem identificadas potenciais RS e extraídos contextos dos respetivos pares de EN. Em segundo lugar, há que desenhar e implementar um algoritmo que consiga, de forma automática, agrupar as RS semelhantes.

[6]https://www.facebook.com/ - Acedido em 03/11/2016

Quanto ao alcance do 2º objetivo, é preciso que se que definam testes que sejam executados em tempo/espaço úteis e que sejam representativos das diferentes possibilidades de *clustering*. Quando se obtiverem resultados, é preciso definir e seguir um método que os avalie.

Este documento apresenta a seguir o trabalho executado para o alcance dos objetivos. No capítulo 2, é feita uma revisão bibliográfica acerca do que está ligado à identificação de relacionamentos em textos. O capítulo 3 apresenta a abordagem que segue o algoritmo desenvolvido, que identifica e agrupa RS entre EN, ao nível concetual e da implementação. No capítulo 4, demonstra-se uma análise exploratória dos dados relacionados com reconhecimento e emparelhamento de EN. Igualmente são demonstrados os resultados dos testes de avaliação (assim como a sua metodologia) à abordagem. E no capítulo 5, são tiradas as conclusões sobre o trabalho de dissertação.

Capítulo 2

Revisão Bibliográfica

A identificação e classificação automáticas de RS entre EN é um processo com vários passos. Começa pelo reconhecimento e localização das EN envolvidas, recolha e processamento dos contextos dessas ocorrências até à aglomeração e identificação dos *clusters* consequentes. Existem soluções relacionadas com cada um dos passos. Há trabalho que lida com todos estes.

Desde já, são descritas as principais características de soluções para a classificação de RS. Hasegawa et al. tiveram uma abordagem sem supervisão, dependente do domínio[3]. Celik et al. criaram a sua solução semi-supervisionada e dependente do domínio[4]. Mais tarde, El Houby apresentou uma abordagem com as mesmas condições que a anteriormente descrita, embora na área da Biologia[2].

Recentemente, Thenmozhi e Aravindan implementaram uma solução que usa PLN, independente do domínio, também para relações taxonómicas e que resolve frases complexas[5]. Esta resolução significa por exemplo a identificação de frases simples que se dispõem ligadas por conjugações.

Ainda existe a solução de Percha et al. que lida com a aprendizagem da estrutura de relações biomédicas a partir de texto não estruturado[14].

No geral, todos os trabalhos tiveram pontuações superiores à metade dos valores máximos possíveis em alguma medida de avaliação. Foi também demonstrada

superação de resultados obtidos em comparação com trabalhos anteriores, nalguns cenários,

2.1 Reconhecimento de EN

El Houby usou um método de reconhecimento de EN baseado num dicionário de termos biológicos como nomes de genes, doenças e proteínas[2]. A constituição do dicionário é baseada em diferentes recursos e bases de dados.

Existe uma solução que permite classificar os tipos de EN, usando uma rede neuronal híbrida[15] que é constituída por dois modelos. Um aprende automaticamente representações de menções, empregando redes neuronais recorrentes para obter representações vetoriais. O outro aprende as palavras dos respetivos contextos, onde *perceptrons* multi-camadas são aplicados para o uso dessa informação, prescindindo assim de características manuais como anotações. O treino da rede foi conseguido com dados gerados através da Dbpedia[1] e Wikipédia. É providenciada uma nova maneira de utilizar a natureza composicional de menções para a tarefa de classificação, o que permite à rede generalizar melhor menções incomuns e não vistas.

Para capacitar o reconhecimento de EN, há também a ferramenta de nome **PAMPO**[1]. Nesta destacam-se duas vantagens:

- Possibilidade de ser integrada numa solução escrita em R[16].
- Identificação de EN na língua portuguesa.

2.2 Obtenção/Manipulação de Contexto

Existe trabalho realizado que lida com contextos processados por intermédio de vetores ou matrizes. Hasegawa et al. lidam com a similaridade dum contexto considerando as palavras que estão entre um par de EN[3]. Cada contexto é representado por um vetor. Este contém os pesos das frequências dos termos que o compõem. Sendo

[1]http://wiki.dbpedia.org/

assim, dados vetores de contextos α e β, a similaridade por cosseno é definida através da seguinte fórmula:

$$\frac{\alpha \cdot \beta}{|\alpha||\beta|}$$

A partir desta operação, é possível determinar o quão semelhantes são dois pares, cujos elementos são pares de EN. O valor desta similaridade varia entre -1 e 1. Quando o valor máximo é atingido, os pares têm exatamente as mesmas palavras de contexto e estas encontram-se na mesma ordem. Se o valor mínimo for atingido, também são exatamente as mesmas palavras de contexto mas em ordens contrárias. A influência da ordem deve-se ao facto as frequências dos termos poderem assumir pesos negativos, se forem mais frequentes entre os elementos dum par de EN na ordem contrária.

Lambrou-Latreille, na sua abordagem que classifica padrões para algoritmos *bootstrapping* de extração de RS, integra algumas medidas de similaridade entre expressões[17].

Turney e Pantel estudaram uma larga gama de aplicações, para processamento semântico de texto, usando três classes de modelos de espaços vetoriais[18]. Os modelos são baseados em matrizes termo-documento, palavra-contexto e par-padrão.

Petroni et al.[19] argumentam que, integrando informação contextual, como metadados sobre fontes de extração, contexto lexical ou informação de tipo, melhora a capacidade de previsão na extração de relações abertas. Estas relações caracterizam-se por formarem conjuntos potencialmente ilimitados de várias fontes tais como bases de conhecimento ou texto em linguagem natural. A abordagem utiliza um novo modelo de fatorização de matrizes.

Contextos podem ser trabalhados com outras estruturas de dados além de vetores e matrizes. El Houby implementou aprendizagem de padrões[2] para a compreensão de contextos. Cada padrão é formado por um par de EN com anotações, indicando os respetivos domínios, e no seu meio estão as palavras que indicam o relacionamento.

Começam por ser analisados os segmentos de texto que ligam pares de sementes que são pares de RS fornecidas por bases de conhecimento. A partir daqui, são agrupadas

substrings semelhantes que ligam esses pares de sementes. Depois são contadas as frequências de cada grupo. Estes passos são repetidos até que nenhum novo grupo seja criado.

A figura 2.1 apresenta as duas primeiras partes que constituem a arquitetura da *framework* proposta. A partir dum corpus, são extraídas as EN, que se encontram nas bases de conhecimento, acabando por ser marcadas. Deste modo, é executado o ciclo de extração dos diferentes contextos. Este termina quando atinge um número especificado de iterações ou não são extraídas mais RS a partir de novas sementes. São iteradas as verificações das frequências dos grupos de distintas RS até ao limiar definido.

Celik et al. utilizaram duas dimensões principais de conceção que influenciam a posterior descoberta de RS: seleção da fonte e esquema de pesagem[4]. As estratégias usadas exploram um grafo para detetar pares de EN que têm um certo tipo de relacionamento num específico período de tempo.

Por outro lado, Thenmozhi e Aravindan fazem uma extração das frases candidatas antes das cláusulas candidatas[5]. Ou seja, antes da obtenção de contextos são verificadas quais as frases (candidatas) a transmitir relações. Depois estas são divididas em cláusulas, onde algumas são as candidatas a conter relações e portanto terão o seu contexto analisado.

Nesta abordagem ainda é aplicada a resolução de conjunções e são tratados relacionamentos inversos. Isto significa que o contexto é processado de modo a ter em conta frases onde EN, como A e B, tenham os mesmos relacionamentos descritos em ambos os sentidos. São exemplo disso as frases:

A equipa ***A*** *venceu a equipa* ***B****.*

A equipa ***B*** *foi vencida pela equipa* ***A****.*

A aplicação da resolução de conjunções serve para lidar com a dificuldade que conjugações em frases impõem no processamento do contexto. Mais detalhes sobre estas tarefas serão dados mais adiante neste capítulo.

Existem outras soluções para processamento de contextos nos quais estão presentes conceitos como EN. Podem então contribuir indiretamente para determinar RS. Jatowt et al. criaram uma solução com a finalidade de emparelhar conceitos semelhantes distanciados temporalmente[20].

Os espaços vetoriais de ambos os conceitos são usados como matrizes, cujas dimensões são o vocabulário e as características de tais vetores representando termos *âncora* (termos frequentes comuns). Esta matrizes servem para ligar os dois espaços temporais.

Também a forma como *perceptrons* multicamadas são aplicados na informação contextual, relativa ao modelo híbrido para a classificação dos tipos de EN, poderia ser adaptada se fossem manipuladas as representações vetoriais obtidas[15].

Há duas propostas que podem ser mais-valias para a simplificação de contextos. Por exemplo, tem-se a frase com 3 EN:

*A equipa **A** triunfou contra a equipa **B**, tal como a **C**.*

Uma simplificação possível da frase seria:

*A equipa **A** venceu a equipa **B**. E a equipa **C** venceu a equipa **B**.*

Desta forma, os contextos relativos aos pares *A*-*B* e *C*-*B* encontram-se parecidos. Assim fica mais explícita a igualdade das RS dos 2 pares.

Na tradução de frases complexas (no entender de pessoas com baixa literacia) para frases simples em português, Specia propõe uma solução[21]. Deste modo, obtêm-se frases léxica e sintaticamente menos difíceis de analisar. O modelo da *framework* baseia-se em probabilidades. Deste surgiu o denominado sistema Moses de tradução máquina estatística que possui 5 funções de recurso.

A experimentação do sistema teve a utilização duma coleção de milhares de artigos noticiosos do Zero Hora e artigos científicos do Folha de São Paulo, e dividiu-se em três

fases (treino, afinação e teste). Foram anotadas frases de referência por um falante nativo de português, sendo executadas 11 operações lexicais/sintáticas distintas.

Simples modificações tinham sido bem conseguidas, enquanto complexas, como inversão de cláusulas, não se concretizaram devido a razões como a ausência de anotações nas palavras. A avaliação efetuada demonstrou alta precisão mas baixo *recall*.

Por outro lado, Gasperin et al. discutem e comparam duas ordens (empírica e hierárquica) de simplificação, lidando com frases adverbiais, para a aplicação dum conjunto de regras durante o desenvolvimento dum sistema para tal propósito[22]. Frases deste género têm considerável impacto na sua complexidade, o que pode influenciar o uso do verbo para a classificação duma RS.

2.3 Identificação/Aglomeração de Relacionamentos

Na descoberta de RS no meio dum par de EN, após o processamento de contextos, são criados *clusters* de pares de EN, adotando o método hierárquico Hasegawa et al.[3] Etapa que é seguida pela etiquetagem de cada *cluster*. As etiquetas caracterizam a relação desses pares e consistem nas palavras em comum. Estas palavras aparecem na maioria dos contextos que pertencem ao *cluster*.

Na figura 2.2, encontra-se uma visão geral não só sobre as duas últimas etapas mas sobre toda a abordagem.

Já a figura 2.3 mostra uma tabela com resultados dos ensaios. Para isto, usaram exemplares de The New York Times do ano 1995, com dois domínios em separado: pessoa / entidade geopolítica e empresa / empresa.

Determinaram três parâmetros para limiares e 5 palavras como tamanho máximo do contexto. Ainda consideraram padrões de expressões paralelas (*e* e *ou* são exemplos) e expressões peculiares do jornal como contexto ignorável. Analisaram o conjunto dos dados manualmente e identificaram as relações para os dois domínios.

Colocam ainda em discussão quatro aspetos do seu método:

- Propriedades das relações: No caso em que são direcionais, conseguir distinguir entre $A \rightarrow B$ e $B \rightarrow A$, e calcular a similaridade de $A \rightarrow B$ com $C \rightarrow D$ e de $A \rightarrow B$ com $D \rightarrow C$. Tal como ocorreu nas experimentações, palavras eram partilhadas por pares de EN alinhadas na direção contrária. Isso levou a erros de aglomeração.

- Quantidade de palavras do contexto: Pares pararam indevidamente em determinados *clusters* na base doutras palavras comuns ocorridas por acidente. Se o número máximo for maior do que a quantidade permitida nas experimentações, poderão ser detetadas palavras comuns adicionais, apesar do ruído aumentar.

- Método de *clustering*: Concluíram que o método de ligação completa é o melhor para aglomerações, comparado com os de ligação singular e ligação média. O limiar da similaridade por cosseno varia na influência sobre os diferentes domínios. Contudo, o método de ligação completa combinado com limite curto de palavras do contexto é útil na descoberta de relacionamentos como demonstraram na avaliação realizada.

- Pares menos frequentes: Descartaram pares de EN onde a coocorrência dos seus elementos é baixa, pois o limiar de frequência era 30. Para tais pares, desde que as variedades dos contextos sejam poucas e as normas dos vetores de contexto sejam demasiado curtas, é difícil de classificar rigorosamente o relacionamento baseado nesses pares. De modo a ultrapassar esta dificuldade, propõem o uso de *bootstrapping*. Com o seu método proposto conseguem resolver como selecionar sementes iniciais.

Antes da descoberta de RS, a solução de Celik et al. processa as postagens, seguindo o desempenho da extração de EN e do enriquecimento semântico[4]. Este passo inicial é um processo que resulta num grafo, ligando os recursos semanticamente enriquecidos, com EN nos recursos correspondentes.

Na figura 2.4 está a arquitetura concetual que expõe ambos os passos, além do processamento inicial e da forma como RS são tipificadas. As estratégias na deteção das RS têm de posicionar por certa ordem pares de EN (de determinados tipos), de modo a que estes, ao estarem *verdadeiramente* relacionados (num específico período

de tempo), apareçam no topo da classificação. A proximidade à veracidade correspondente aos relacionamentos *verdadeiros* é obtida via DBpedia e um estudo de utilizador.

A figura 2.5 demonstra a precisão perante as duas configurações de teste, diferenciando as estratégias baseadas em *tweet*+notícia, em notícia e em *tweet*. A *MRR* (classificação recíproca média) indica em qual posição o primeiro relacionamento *verdadeiro* ocorre em média. Já $P@20$, $P@10$ e $P@5$ (precisão nos melhores 20, 10 e 5) representam a proporção média de relacionamentos *verdadeiros* dentro dos *tops* 20, 10 e 5 respetivamente.

Por exemplo, RS entre um par de EN, onde os seus elementos tinham os tipos *pessoa* e *evento*, foram fáceis de detetar. O contrário não se verifica nas RS entre elementos do tipo *pessoa*. A diferença é vista na figura 2.6.

Para RS com menor período de tempo, estratégias baseadas no Twitter são mais precisas na sua deteção. Já estratégias baseadas em notícias são melhores para detetar RS que perduraram. O gráfico esquerdo da figura 2.7 elucida esse facto. Na avaliação das estratégias, usaram mais de 10 milhões de *tweets* e 70 mil artigos de notícias. A *framework* permite uma base para a melhoria na pesquisa e exploração de conteúdos, assim como na recomendação dos mesmos.

Especificamente em Biologia, há soluções para a extração de RS a ter em conta. A solução de El Houby parte do pré-processamento do corpus (com 1000 resumos das coleções de MEDLINE, OMIM e UNIPORT) feito por um *named entity recognizer* que a própria implementou (como foi dito anteriormente)[2].

O algoritmo de extração de relações biológicas utiliza reconhecimento de padrões com vista à coleção de bases de conhecimento. Todo o procedimento está presente na figura 2.8. Gerados os padrões dos vários documentos que relacionam termos biológicos diferentes, são usados para a extração dos pares relacionados. O processo é iterado até chegar a conhecimento extraído.

Como resultados experimentais, extraiu pares entre doenças e genes usando padrões como *causada por* e *associado com*. Ainda outros tipos de relações tinham sido

extraídos, como os que estão na tabela da figura 2.9.

Num domínio relacionado, Percha et al. descrevem um novo algoritmo, denominado *Ensemble Biclustering for Classification* (EBC)[14]. Este consiste na aprendizagem da estrutura de relacionamentos biomédicos automaticamente a partir de texto, sobressaindo diferenças na escolha de palavras e na estrutura das frases.

Validam a performance do EBC partindo desde conjuntos curados manualmente de relacionamentos farmacogenómicos de PharmGKB e de relacionamentos *droga-alvo* de DrugBank. Depois aplicam EBC para mapear o completo universo de relacionamentos *droga-gene* baseados nas suas descrições em Medline. O resultado da aplicação revelou uma estrutura inesperada que desafia noções correntes acerca de como esses relacionamentos são expressados em texto. Usam também o EBC para descobrir novos relacionamentos *droga-gene* para ambas as bases de conhecimento.

PLN é um caminho para a extração de RS. Prova disto é o processo de aprendizagem de relações, na *framework* de Thenmozhi e Aravindan[5], que implica a extração de frases candidatas (dados conceitos) e a sugestão de etiquetas de relações. Entre estas duas tarefas, são identificadas as cláusulas e resolvidas as conjunções. Usam-se regras independentes do domínio para a geração das cláusulas.

De seguida, identificam-se triplos candidatos onde se definem as relações. Os três elementos que definem uma relação são a etiqueta que representa esta mesma e um par de EN. Sugestões de etiquetas realizam-se usando uma base de dados lexical (Wordnet[23]) e a frequência das relações já obtidas. O esquema com as etapas da *framework* está na figura 2.10.

Nas experimentações, trabalharam sobre duas coleções de documentos: uma no domínio da agricultura e outra no domínio dos computadores. A partir destes, criaram-se duas ontologias que são apresentadas nas figuras 2.11 e 2.12.

A avaliação aos diferentes resultados demonstraram que o seu método foi sempre melhor comparativamente a outros.

Ainda a propósito de PLN, Silva et al. anunciam o primeiro *parser* robusto amplamente disponível para português[24].

No que toca também a RS complexas, Mesquita propôs métodos de extração de relações abertas baseados em regras que alcançam alta efetividade, num custo computacional baixo em comparação com abordagens anteriores[25]. O problema foi endereçado da extração de relações aninhadas (aceitam instâncias de relações como argumentos) e relações n-árias ($n > 2$ argumentos).

Ainda descreveu uma solução elegante que começa com métodos de extração superficiais, e que decide dinamicamente e numa base por frase, se desenvolve ou não métodos de extração mais profundos baseados em análise de dependências e etiquetagem de papéis semânticos. A solução prioriza recursos computacionais extraordinários para frases, descrevendo instâncias de relações que estão favoravelmente a ser extraídas pelos métodos mais profundos.

Com foco no processo de aglomeração, não apenas de RS, há soluções interessantes. Uma abordagem automática, dependente do domínio, para identificar RS semelhantes, foi trabalhada por Wang et al.[26] A aglomeração sucede-se a partir de um grafo tripartido com constrangimentos. A informação de tipos usada pode servir como uma supervisão indireta para a aglomeração. De modo a se verificar a supervisão, são derivados os constrangimentos ou da proximidade à veracidade de entidades e relações ou baseados em conhecimento automaticamente induzido a partir dos dados.

As relações aglomeram-se a partir da extração de informação aberta diretamente baseadas em estatísticas de dados e apenas usando EN como constrangimentos. Estas EN possibilitam a formação de pares *must-link* e *cannot-link*. No primeiro caso, as EN devem encontrar-se num mesmo *cluster*. Enquanto no segundo caso, devem encontrar-se em *clusters* distintos. Este pares têm o propósito de melhorar os resultados de *clustering*.

O algoritmo tem uma função-objetivo que se pretende minimizar. Esta procura otimizar a atribuição de etiquetas a *clusters*. O mesmo algoritmo pode também ser aplicado em bases de conhecimento para normalizar diferentes relações com *clusters* atribuídos, especialmente relações *multi-hop* que são conjugações de múltiplas relações.

As experimentações concretizaram-se com o uso da base de conhecimento anotada humanamente Freebase e de dados extraídos por ReVerb (que é um sistema de extração de informação aberto).

Avaliaram através de informação mútua normalizada, com efeitos de constrangimentos de entidades e de relações. Esta medida pontua 1 se os resultados de aglomeração correspondem às etiquetas da categoria perfeitamente e 0 se os *clusters* são obtidos duma partição aleatória.

O respetivo método superou os outros cinco em ambos os cenários em que houve comparação.

Ainda relativo a processos de *clustering* que usam pares *must-link* e *cannot-link*, há a destacar a abordagem semi-supervisionada de Basu et al. que aplica o campo aleatório de Markov[27]. O problema é formulado através da minimização, por uma função-objetivo, onde cada elemento (além da parte da distância a cada centróide) possui uma parte relativa ao custo desse elemento pertencer ou não a cada cluster.

São seguidos dois algoritmos para constrangimento de pares: *Explore* e *Consolidate*. O primeiro essencialmente usa o esquema *farthest-first* para formar consultas apropriadas de modo a obter os vizinhos disjuntos em termos de emparelhamento requerido. O segundo serve para consolidar a estrutura obtida na fase anterior.

Esta abordagem enquadra-se num caso de aprendizagem ativa, que consiste em consultar interativamente uma fonte de informação. A finalidade deste tipo de aprendizagem é obter os outputs desejados nos novos pontos de dados.

Na avaliação, o k-Means de aglomeração constrangida por emparelhamento ativo provou que ambos os algoritmos são importantes e superiores (quando combinados) para a obtenção de melhores resultados em comparação com o k-Means de aprendizagem não supervisionado. Concluíram ainda que não existe para já ruído nos constrangimentos da *framework* atual.

2.4 Sumário

Reviu-se trabalho relacionado com vários processos que levam à identificação e aglomeração de RS. Inclusive, demonstrou-se diferentes perspetivas para tarefas idênticas. O estudo até aqui realizado permite definir um ponto de partida para a criação duma abordagem. A solução a apresentar deve tanto reconhecer como emparelhar EN. Finalmente, deve desempenhar *clustering* de pares com base nos seus contextos a extrair. No capítulo 3, será então apresentado um algoritmo não supervisionado no sentido em que aglomera pares de EN com base no contexto sem constrangimentos de qualquer fonte de informação.

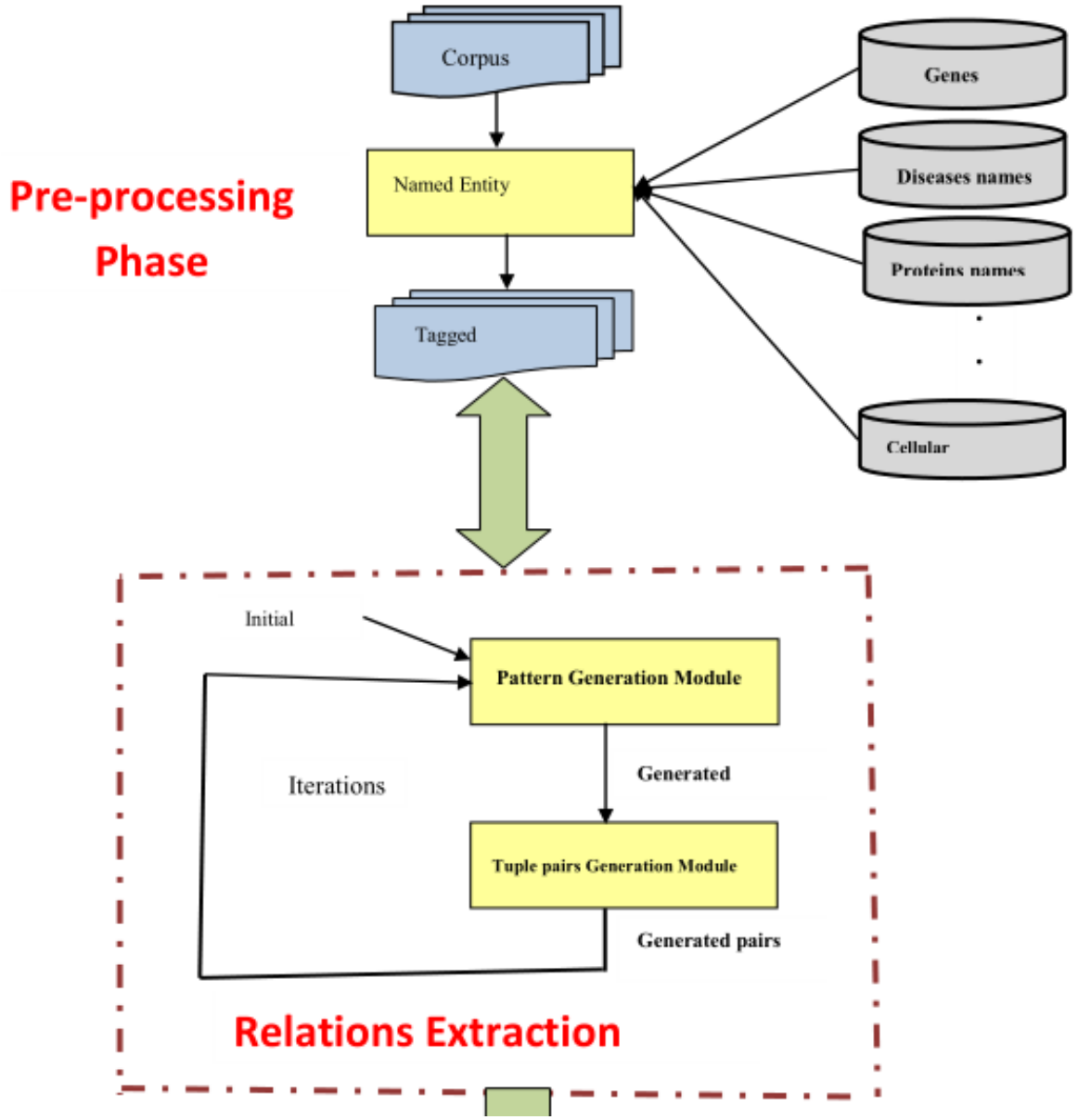


Figura 2.1: Extração de RS, usando pré-processamento, com um módulo que gera padrões a partir de cerca de 5 pares válidos. Estes funcionam como sementes, sem qualquer treino e servem para a geração de pares na forma de tuplos.[2]

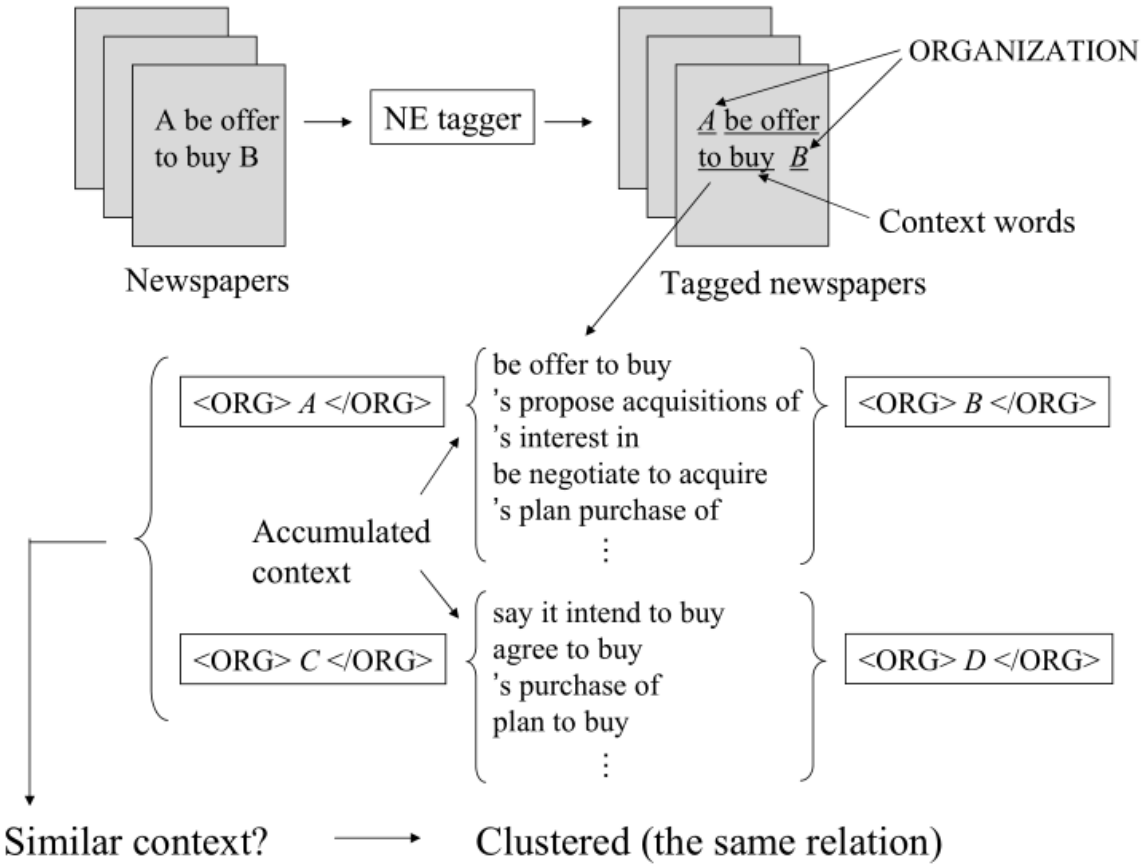


Figura 2.2: Ideia básica de Hasegawa et al.[3] que começa pela colocação de *tags* nas EN, identificando os seus tipos, e pela localização dos contextos. Quando o contexto acumulado dum par é semelhante ao de outro, terão a mesma relação à partida e por isso são aglomerados.

Major relations	Ratio	Common words (Relative frequency)
President	17 / 23	President (1.0), president (0.415), ...
Senator	19 / 21	Sen. (1.0), Republican (0.214), Democrat (0.133), republican (0.133), ...
Prime Minister	15 / 16	Minister (1.0), minister (0.875), Prime (0.875), prime (0.758), ...
Governor	15 / 16	Gov. (1.0), governor (0.458), Governor (0.3), ...
Secretary	6 / 7	Secretary (1.0), secretary (0.143), ...
Republican	5 / 6	Rep. (1.0), Republican (0.667), ...
Coach	5 / 5	coach (1.0), ...
M&A	10 / 11	buy (1.0), bid (0.382), offer (0.273), purchase (0.273), ...
M&A	9 / 9	acquire (1.0), acquisition (0.583), buy (0.583), agree (0.417), ...
Parent	7 / 7	parent (1.0), unit (0.476), own (0.143), ...
Alliance	3 / 4	join (1.0)

Figura 2.3: Tabela que demonstra os maiores *clusters*, por cada domínio, juntamente com o rácio do número de pares correspondendo à maior relação para o total número de pares em cada *cluster*. Do lado direito, as palavras comuns mais frequentes e as suas frequências relativas. Estes dados resultam das experimentações de Hasegawa et al.[3]

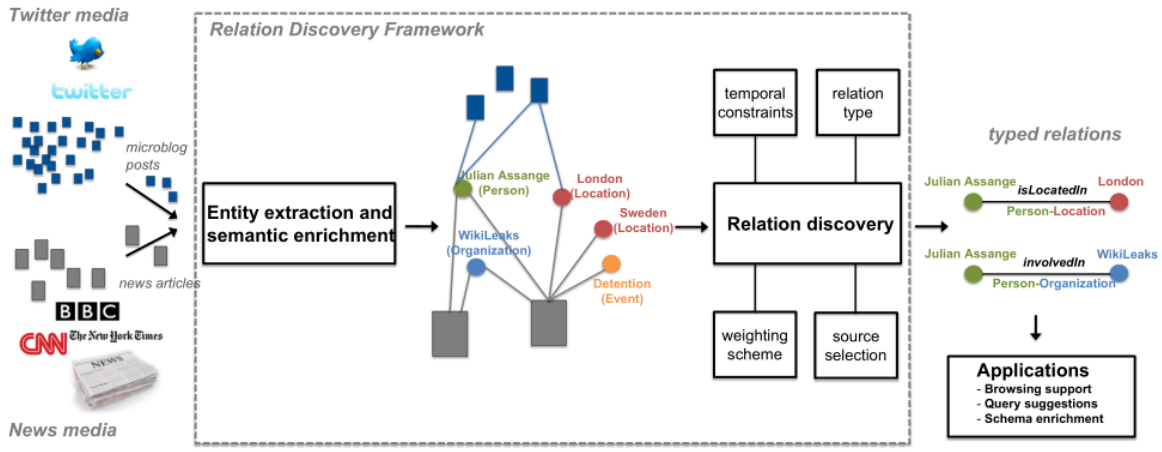
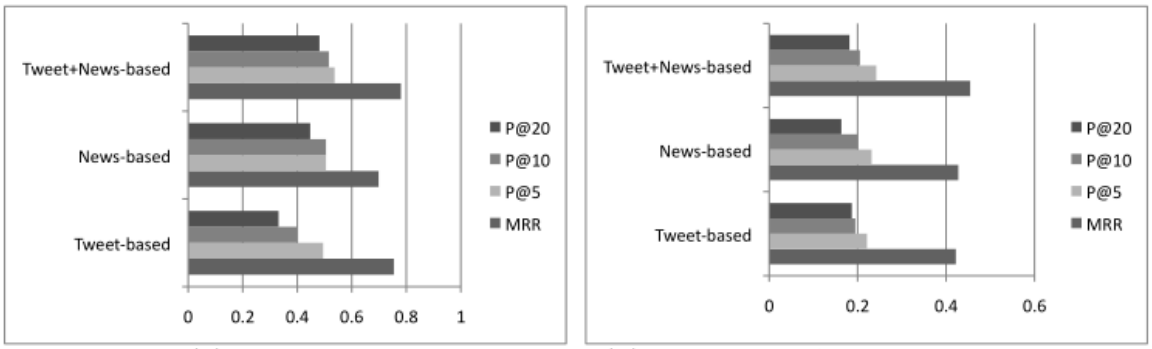
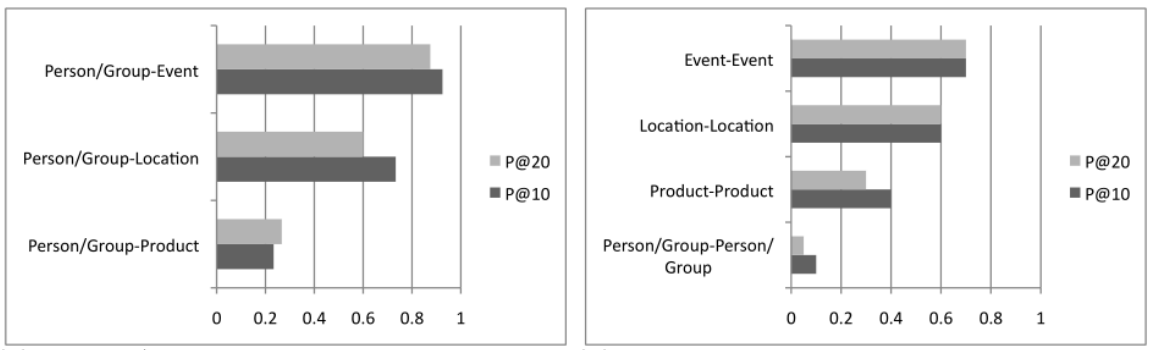


Figura 2.4: Arquitetura concetual da *framework* de descoberta de RS baseadas no Twitter.[4] São recolhidos *posts* do *microblog* e artigos de notícias. Procede-se então ao 1º principal passo (extração de EN e enriquecimento semântico). Gerado o grafo, onde as EN e os respetivos tipos estão representados nos nós, é dado o 2º e último passo (descoberta de RS tendo em conta seleção da fonte, esquema de pesagem, constrangimentos temporais e tipo de relação). Por fim, saem as relações tipificadas que podem servir para três aplicações.



(a) user study (b) prediction of DBpedia relationships

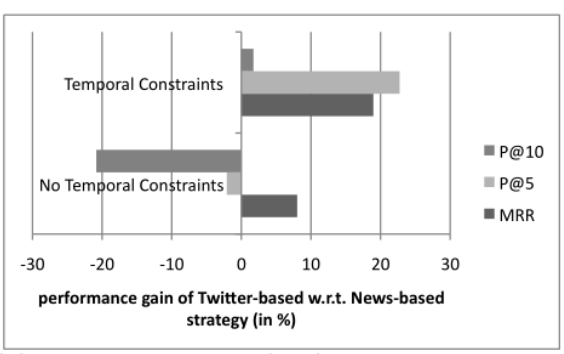
Figura 2.5: Precisão das diferentes estratégias de descoberta de RS baseadas em (a) a proximidade à veracidade obtida do estudo de utilizador e (b) a proximidade à veracidade obtida da DBpedia (sendo mais que 5000 RS).[4]



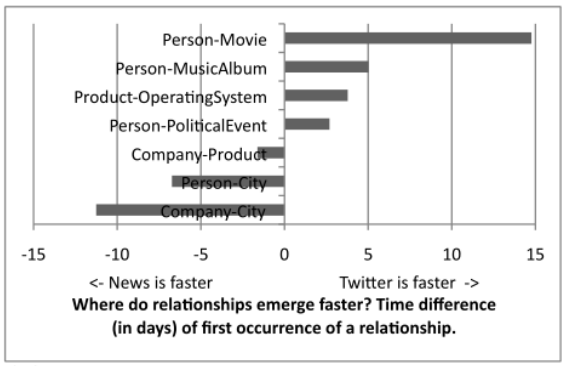
(a) Person/Group relations to events, locations and products.

(b) Relations between entities that have the same type.

Figura 2.6: Precisão dos diferentes tipos de RS: (a) de pessoas ou grupos com eventos, locais e produtos; (b) entre EN que são do mesmo tipo.[4]



(a) Relations with(out) temporal constraints.



(b) Time difference in detecting relations.

Figura 2.7: Aspetos temporais: (a) diferença na precisão entre as estratégias baseadas em notícias e baseadas em *tweets* para RS com/sem constrangimentos temporais; (b) diferença de tempo entre as estratégias baseadas em notícias e baseadas em *tweets* na deteção de certas RS.[4]

1. Use N.E.R. to tag biological terms in the corpus
2. **Do** to collect different relations types for different KBs
3. Provide initial seeds of relation pairs according to target KB
4. **Repeat**
5. **Repeat**
6. Analyze the segments of text that connects seeds pairs
7. Group similar substrings that connect seeds pairs
8. Count the frequency of each group
9. **Until** no new group can be generated
10. **For** each created group
11. **If** group frequency > threshold
Then generate pattern
12. Rank generated patterns by frequencies
13. **Repeat**
14. Use generated patterns to extract relation pairs
15. Compare the extracted relation pairs with those in the target KB
16. **If** No similar relation's pairs in KB
Then add new pairs to the target KB
17. **Until** no new pairs can be extracted
18. Use extracted relation pairs as seeds
19. **Until** termination condition
20. **End Do**

Figura 2.8: O processo completo para a extração de RS e coleção de várias bases de conhecimento[2].

Target Knowledge base	Pattern (Relation Type)	Extracted Pairs
Protein-Protein	Interact with	(Battenin, BIP)
Gene-Protein	Encode	(RAC2, GTP-binding)
Protein-Disease	Cause	(Presenilin-2, Alzheimer)
Gene-Protein	Encode	(LBR, bifunctional protein)
Gene-Disease	Affected with	(NKX2.5, congenital heart disease)
Gene-Disease	in	(NKX2-5, ventricular septal defects)
Protein-Protein	interact with	(heterogeneous nuclear ribonucleoprotein, TAR DNA-binding protein 43)

Figura 2.9: Amostra de padrões gerados e pares de RS[2].

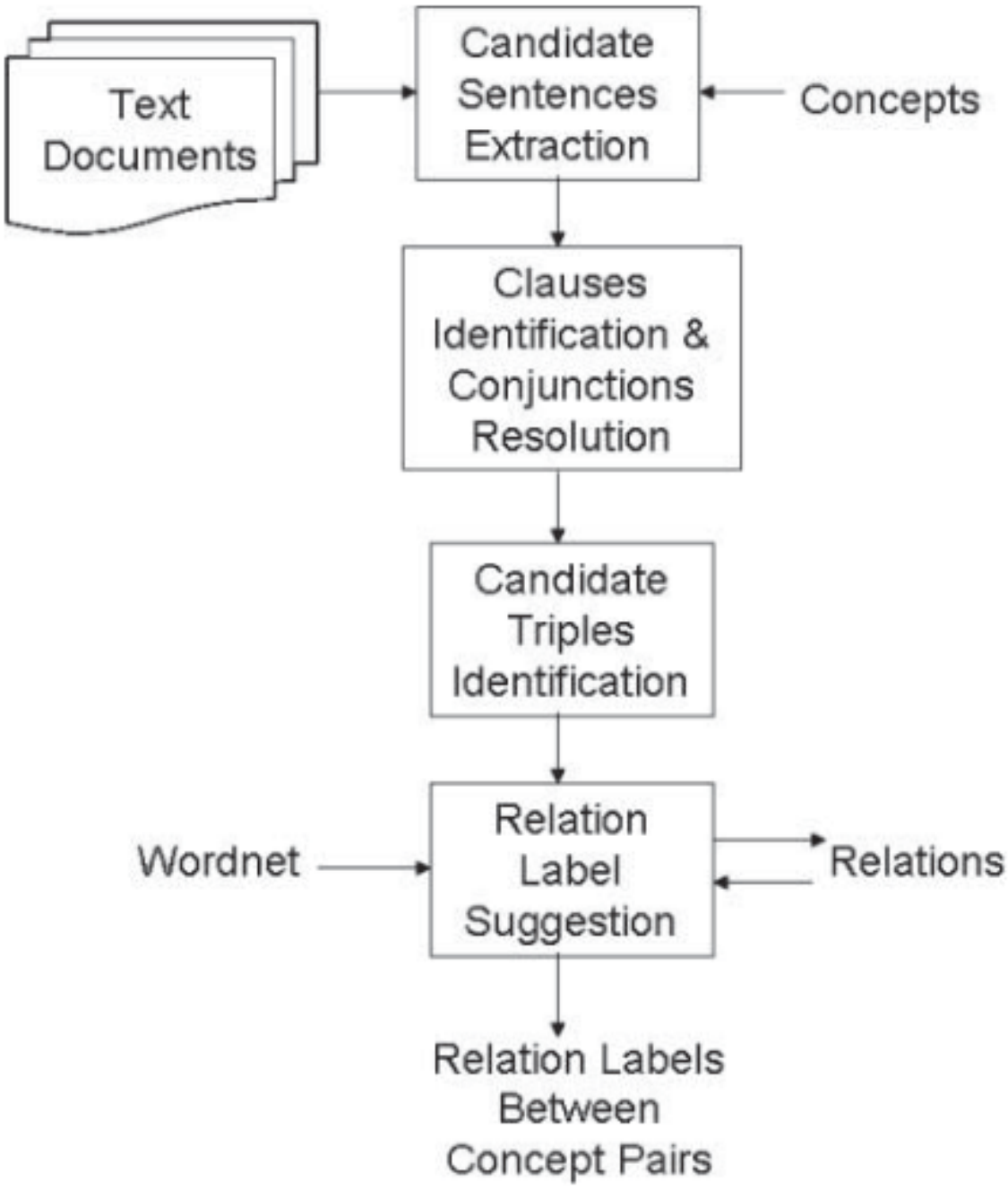


Figura 2.10: *Framework* proposta por Thenmozhi e Aravindan[5].

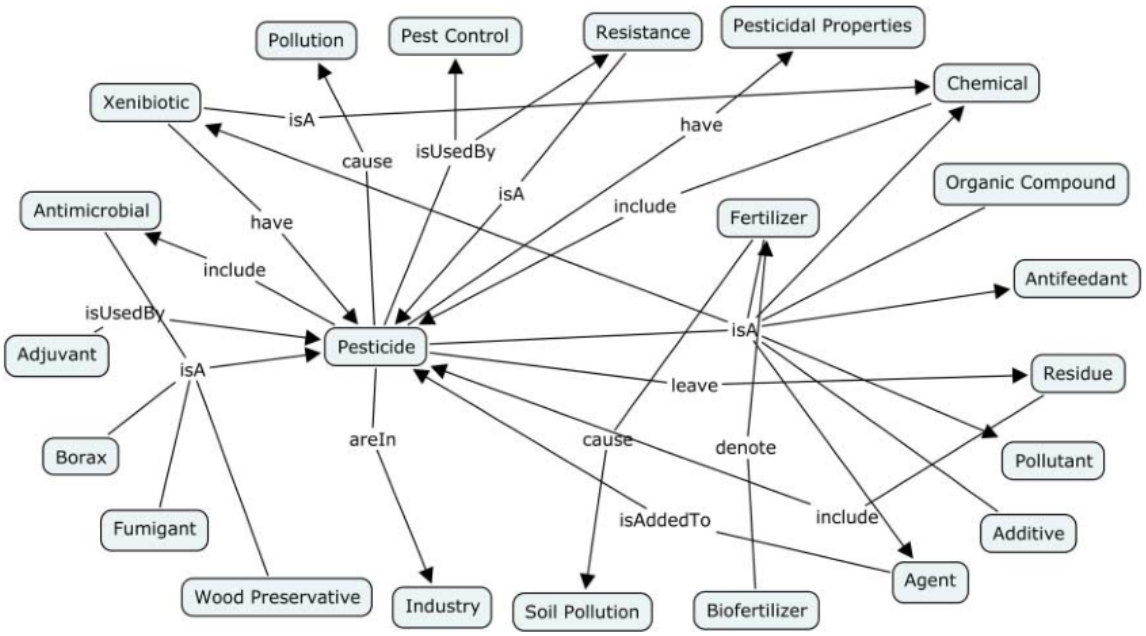


Figura 2.11: Ontologia para agricultura criada a partir do método de Thenmozhi e Aravindan[5].

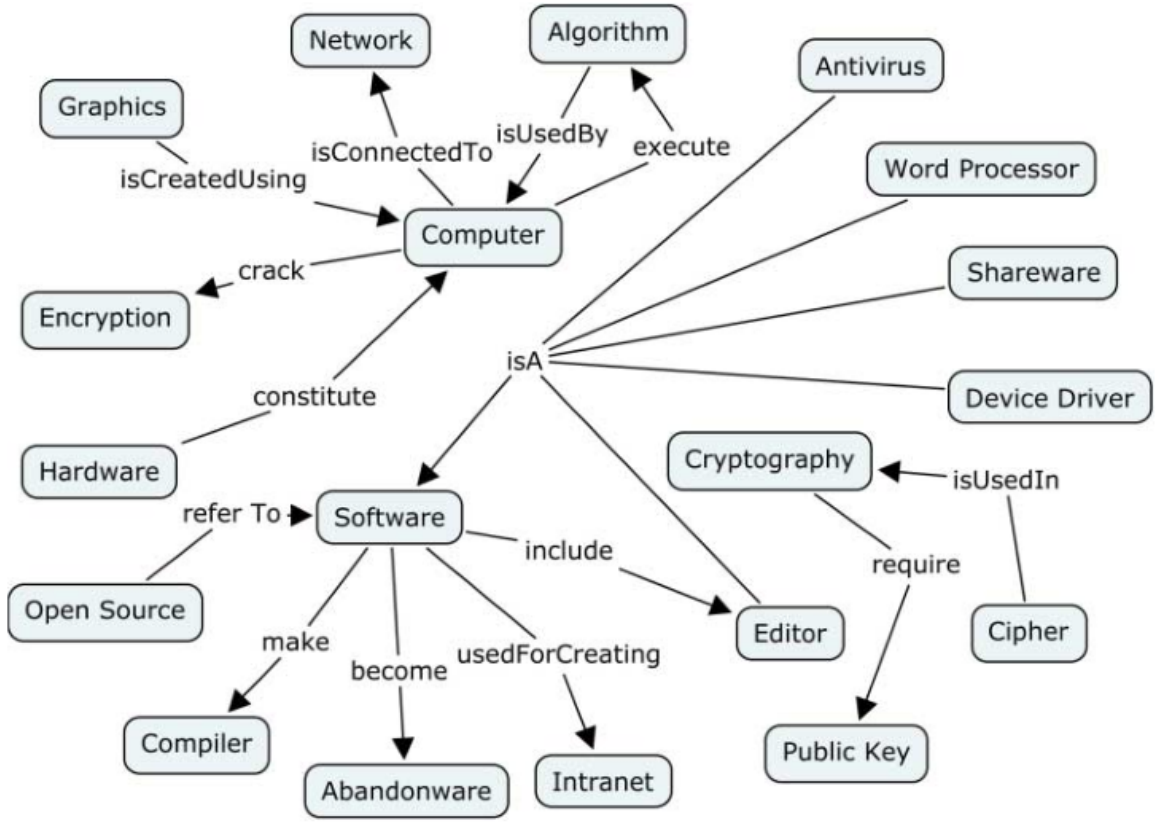


Figura 2.12: Ontologia para computadores criada a partir do método de Thenmozhi e Aravindan[5].

Capítulo 3

Abordagem

A abordagem que se propõe aglomera RS entre pares de EN de forma não supervisionada e é independente do domínio no que toca às EN e aos contextos. A não supervisão justifica-se, além da inclusão da análise de *clustering*, por nomeadamente não haver constrangimentos no emparelhamento de EN nem na aglomeração dos pares. Toda a solução foi escrita na linguagem R[16]. Na primeira fase, faz o reconhecimento automático das EN, usando **PAMPO**[1]. Este é apropriado para a deteção e localização de EN num texto, num ou mais ficheiros no formato TXT. No anexo A, encontram-se detalhes sobre a execução da sua função principal.

Nas fases posteriores da abordagem, são usadas funcionalidades do pacote **tm**[28] pela vantagem na execução de tarefas de *text mining*. O resto da solução chama funções básicas de R. A figura 3.1 apresenta uma visão geral desde o reconhecimento das EN até ao processo de aglomeração de RS. Cada retângulo representa dados, onde o primeiro é relativo ao(s) texto(s) e os seguintes serão relativos a quadros de dados resultantes de processos (representados pelos objetos com fundo verde).

Havendo um conjunto de textos, não estruturados e sem anotações, é efetuado o reconhecimento das EN. Daqui obtém-se cada EN tal como se encontra no texto, a sua desambiguação e o seu posicionamento. Podem ser escolhidas EN a retirar e/ou eliminados registos específicos do quadro de dados que diz respeito a estas.

Depois são detetadas as possíveis RS entre pares de EN e extraídos os seus contextos através das localizações duma das EN que forma cada par. Consoante o que for

previamente parametrizado, extraem-se ou as palavras no meio dos elementos ou todas as palavras da frase onde os elementos se inserem. Desta forma, é permitida a exploração de ambos os cenários com outros parâmetros que influenciam posteriormente a aglomeração.

Obtidos os pares e os respetivos contextos, estes são agregados num só contexto tendo em comum o mesmo par. Isto é feito para que existam pares únicos e para que todos os contextos dum par sejam caracterizados por completo.

O resultado da agregação são registos em que cada um contém um par distinto de EN e uma lista com todos os seus contextos. Por fim, é aplicada técnica de *clustering* dos documentos (como são vistas as listas de contextos). Daqui obtém-se um quadro com os elementos que formam os pares de EN, com os números dos *clusters* a que foram atribuídos e as etiquetas que classificam as RS.

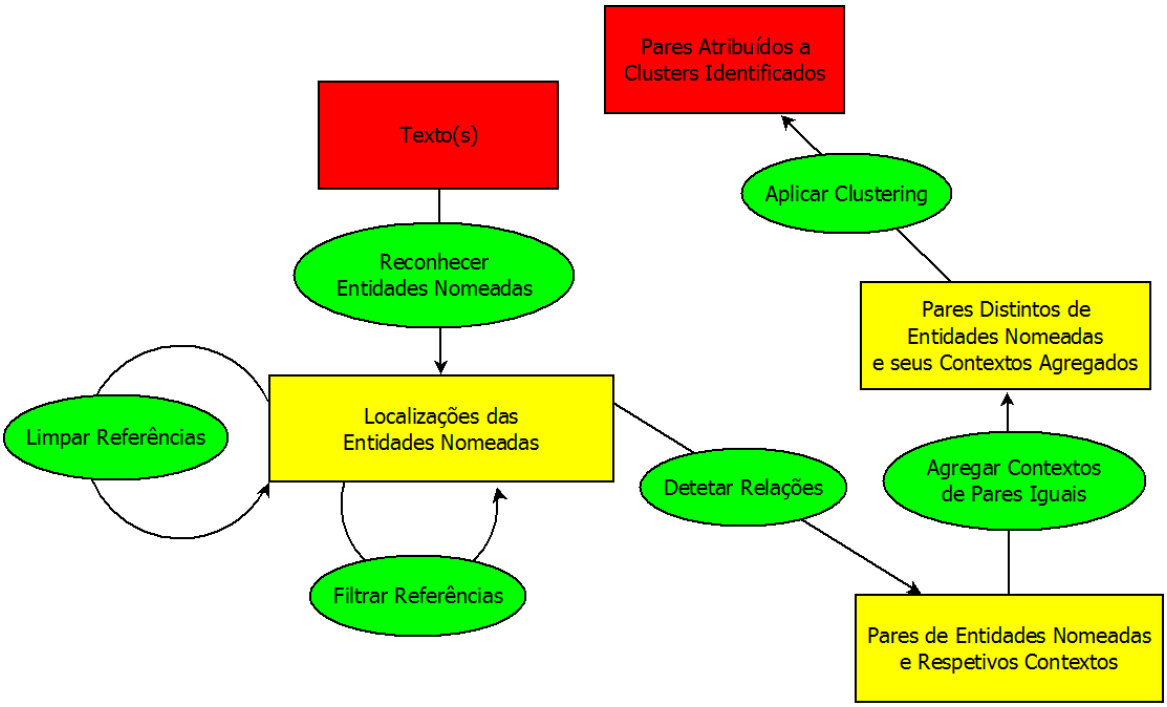


Figura 3.1: Abordagem proposta para agrupar RS a partir dum conjunto de textos não estruturados.

Nas secções restantes deste capítulo, são explicadas as várias etapas da abordagem proposta. Para cada etapa, haverá uma componente concetual e outra relativa à

sua implementação. Concetualmente, são analisados as dificuldades para a solução de cada etapa. Os aspetos da implementação incluem informação sobre quadros de dados, resultantes do processamento concretizado na respetiva etapa.

Na secção 3.1, é explicado o reconhecimento das EN, tendo em conta as suas localizações e desambiguações. A secção 3.2 introduz uma forma de detetar EN que partilhem a mesma frase e depois o seu emparelhamento de tal forma que potenciais RS sejam adquiridas. Esta etapa envolve a sub-etapa da secção 3.3. Esta secção apresenta o que é concretizado para os termos do contexto, tanto intermédio como completo, serem aproveitados para a comparação entre pares de EN. E na secção 3.4 é descrita a etapa onde é definido o *clustering*. Também aqui se descrevem vários desafios desde a transformação das listas de contextos em documentos até à etiquetagem dos *clusters*.

3.1 Tratamento das EN Reconhecidas

Esta primeira etapa compõe a fase do pré-processamento. Desde a introdução de informação sobre os textos onde se pretende que haja reconhecimento das EN até que os dados necessários destas se organizem, é necessário ultrapassar alguns desafios:

- Absorver o conteúdo dum texto t ou a localização dum diretório *dir* onde existem textos.

- Usar uma função que, ao analisar o conjunto de textos, consiga corresponder certas sequencias de termos a EN.

- Situar cada EN, através da sua localização, ao nível do documento d, parágrafo p e frase f. De seguida, confirma-se que a *string* da EN está contida na *string* da frase. A sua importância deve-se ao facto de possibilitar saber que EN partilham uma frase e portanto saber quais destas podem emparelhar. Também a localização das EN é útil para a extração dos seus contextos.

- Perceber quando uma mesma EN representa conceitos diferentes. Ou seja, saber se m é uma menção da entidade e_1, e_2, etc. Perante EN iguais, é necessário que haja desambiguação. Um dos casos onde há utilidade na distinção é quando EN são abreviaturas. Por exemplo, saber se a menção APN diz respeito à Associação

Portuguesa dos Nutricionistas[1], à Associação Portuguesa de Neuromusculares[2] ou a outros conceitos. Deste modo, evita-se o emparelhamento de diferentes EN que são relativas ao mesmo conceito, pois pretende-se que os elementos dum par sejam distintos.

Inicialmente a implementação da abordagem permite a entrada de textos de forma a que nestes se reconheçam e localizem EN. Estas tarefas são desempenhadas por **PAMPO**[1]. Caso o parâmetro específico indique que o input é uma pasta com ficheiros, o resultado é um quadro de dados com 5 atributos: *ficheiro*, *parágrafo*, *frase*, *EN* e *desambiguação de EN*. Caso o mesmo parâmetro indique que é um texto, o quadro resultante não possui a 1ª coluna. A tabela 3.1 explica o conteúdo de cada coluna.

Atributo	**Descrição**
File	Nome do ficheiro.
Paragraph	n-ésimo parágrafo.
Sentence	n-ésima frase.
Entity	EN tal como se encontra no texto.
Entity_desamb	Desambiguação da EN com nome distintivo.

Tabela 3.1: Atributos do quadro de dados, gerado pela **PAMPO_pt** que localiza as EN em textos de ficheiros, na língua portuguesa.[1]

Os dados pertencentes à 1ª, 4ª e 5ª coluna do quadro encontram-se como *string*. Ainda há possibilidade de remover registos específicos de EN reconhecidas que não devem continuar no processo. Outra possibilidade é a de definir um conjunto de exceções de EN que não são consideradas independentemente da sua quantidade e donde ocorrem.

Por fim, enquanto a remoção das localizações de EN únicas nas suas frases é a opção por defeito, existe a possibilidade de dispensar EN que contribuem para outras frequências em frases. O que é favorável quando se pretende trabalhar, por exemplo, com frases que tenham apenas duas EN.

[1] http://www.apn.org.pt/
[2] http://apn.pt/apn/

3.2 Localização de RS

A segunda etapa corresponde também ao pré-processamento e tem a ver com o foco em potenciais RS. São identificadas algumas dificuldades relativas ao emparelhamento de EN:

- Saber quais EN estão numa mesma frase. Ou seja, agrupar EN que partilhem o documento d, o parágrafo p e a frase f.
- Emparelhar EN, distinguidas inclusive pela desambiguação, de acordo com dois critérios. O primeiro é a opção de se formarem pares, cujos elementos ou se encontram apenas consecutivamente na mesma frase ou combinam de todas as maneiras numa frase. O outro critério tem a ver com determinados pares $a - b$ e $b - a$ serem considerados o mesmo par. Quando este critério está definido desta maneira, são ordenados lexicograficamente os elementos dos pares que não estejam nesta situação. Por outro lado, quando o elemento a aparece à esquerda do elemento b é formado o par $a - b$ e forma-se o par $b - a$ se o contrário se suceder.
- Situar os contextos de cada par de EN. Significa que, sabendo do documento d onde ambos os elementos dum par se encontram, é preciso extrair a frase correta a partir do parágrafo p e depois a partir da frase f.

Dadas as EN reconhecidas pré-processadas, passa a ser permitida a localização de possíveis RS. Existe o parâmetro *only_consecutive* que, quando verdadeiro, limita a criação de pares com EN consecutivas. Para a tarefa de localização, foi implementada em primeiro lugar a função **scanIteratively** que usa um ou dois ciclos iterativos para verificar que pares de EN poderão ter um relacionamento. Em segundo lugar, definiu-se a função **scanEfficiently**. Esta caracteriza-se por inicialmente selecionar cada conjunto de registos que tenham o ficheiro, o parágrafo e a frase em comum.

Na **scanIteratively**, o primeiro ciclo percorre todos os registos das EN verificando, em cada iteração, se cada uma pode formar par com a do registo sucessor. Já o segundo percorre os registos a seguir ao sucessor, que demonstrem que a respetiva EN segue uma condição juntamente com a do registo correspondente à iteração do 1º ciclo e se $only\_consecutive = FALSE$. Aqui é verificada a possibilidade de formação de pares. O ponto de vista assintótico deste método pode ser visto na secção 3.5. Com

o objetivo de validar a paridade de EN, é preciso antes confirmar se ambas partilham a mesma frase e se as respetivas desambiguações são diferentes.

Na **scanEfficiently**, selecionados todos os registos relativos a cada frase, é analisado um destes, a cada iteração, sendo processado com o posterior caso $only\_consecutive = TRUE$. Se a condição não se verifica, os restantes registos a seguir também são processados um a um com o registo da iteração. Assintoticamente, a função é averiguada na secção 3.5. A validação dum par necessita unicamente da distinção das desambiguações nos respetivos elementos.

Ambas as funções retornam resultados iguais, embora a segunda seja mais modular. Ainda há a ter em consideração que depois de formados os pares e extraídos os seus contextos, o parâmetro $order.significance$ quando toma valor verdadeiro obriga à ordenação dos elementos dos pares por qual aparecem nas respetivas frases. A tabela 3.2 mostra que atributos resultam destes métodos.

Atributo	**Descrição**
entity1.name	Elemento de EN numa frase de texto processado.
entity2.name	Elemento distinto de EN na mesma frase de texto processado.
context	Contexto extraído onde estão presentes ambas as EN.

Tabela 3.2: Pares de EN resultantes da localização de potenciais RS.

3.3 Extração de Contextos

A etapa, que se insere na de localização de pares para RS, é a da obtenção do contexto relativo a esses pares. O contexto neste caso é visto como um vetor de palavras, ignorando outros símbolos como sinais de pontuação ou números, da seguinte forma:

$$C = C_{left}, NE_{left}, C_{middle}, NE_{right}, C_{right}$$

.

Focando em cada componente do contexto, entende-se o seguinte:

- $C_{left} = \ldots, w_{left_h}, \ldots$ é a parte à esquerda do par de EN.
- $NE_{left} = \ldots, w_{1^{st}NE_i}, \ldots$ termos que definem a EN do lado esquerdo.
- $C_{middle} = \ldots, w_{middle_j}, \ldots$ é a parte entre os elementos do par de EN.
- $NE_{right} = \ldots, w_{2^{nd}NE_k}, \ldots$ termos que definem a EN do lado direito.
- $C_{right} = \ldots, w_{right_l}, \ldots$ é a parte à direita do par de EN.

Deste modo, w_{left_h}, $w_{1^{st}NE_i}$, w_{middle_j}, $w_{2^{nd}NE_k}$ e w_{right_l} são as palavras número $h/i/j/k/l$ das respetivas componentes. Além disso, $0 \leq h \leq m, 0 \leq i \leq n, 0 \leq j \leq o, 0 \leq k \leq p, 0 \leq l \leq q$, onde m, n, o, p e q são as quantidades de palavras de cada uma das cinco componentes.

Posto isto, é necessário ultrapassar:

- A extração dos termos certos do contexto que diz respeito ao par.
- A atribuição correta do contexto extraído a cada par formado.
- A agregação a seguir de todos os contextos num só, que partilhem o mesmo par.

Portanto, uma opção importante na implementação, para a extração contextual, é *what_context* que tem um de dois valores: *between* e *all*. Após a extração, independentemente da atribuição a *what_context*, são retirados os algarismos e outros símbolos dos contextos.

O primeiro valor diz respeito à extração da *string* que pode ser vista como o vetor C_{middle}. A escolha exclusiva de C_{middle} já fez parte de trabalhos semelhantes[3, 2]. Num destes, a escolha foi justificada com a existência de menor ruído em relação à escolha de todas as palavras[3]. Além de que El Houby indica que o contexto intermédio, em documentos na língua inglesa, é considerado como o mais indicativo de relacionamentos entre termos[2].

Se se optar pela extração completa, todas as palavras são aproveitadas. Numa fase posterior, NE_{left}, NE_{right} e outras EN reconhecidas que formem ou façam parte de C_{left}, C_{middle} e C_{right} não permanecerão. Um trabalho com um processo deste tipo usa toda a frase, embora através de PLN[5].

Outras operações sobre contextos serão feitas à posteriori, entre as quais, que deem origem a sacos-de-palavras tal como acontece na solução de Hasegawa et al.[3]. Antes disso, são agregados os contextos que nos seus registos partilhem as mesmas EN pela mesma ordem. O resultado é um quadro com a mesma dimensão horizontal, mas em princípio com menos instâncias e diferente no tipo de dados da 3ª variável. A tabela 3.3 esclarece o seu conteúdo.

Atributo	**Descrição**
entity1.name	Elemento de EN numa frase de texto processado.
entity2.name	Elemento distinto de EN na mesma frase de texto processado.
context	Lista de todos os contextos onde estão presentes ambas as EN.

Tabela 3.3: Pares de EN resultantes da localização de potenciais RS, com os seus contextos agregados.

3.4 *Clustering* das RS

Quando obtidos os pares de EN com as respetivas listas de contextos, chega-se então à etapa onde são aglomerados pares com semelhanças ao nível das RS. Até que se encontrem pares semelhantes, é necessário o desempenho de tarefas que levantam algumas objeções:

- Considerar cada lista de contextos dum par como um único documento.
- Constituir um corpus a partir dos documentos.
- Saber quais os mapeamentos a aplicar nesse corpus.
- Procurar forma de quantificar o peso das frequências dos termos nos documentos.
- Medir o grau de semelhança entre os documentos com base nos seus termos.
- Determinar qual o número indicado de *clusters* a obter.
- Escolher um método de *clustering* para esses documentos.
- Corresponder os *clusters* aos pares de acordo com os respetivos documentos.
- Caracterizar semanticamente cada *cluster* com termos usados como etiquetas.

Na implementação da derradeira etapa, é definida a tarefa de *clustering*. O objeto usado para tal é um corpus com documentos que dizem respeito aos contextos dos pares formados. Portanto, cada lista de contextos passa a ser processada como um documento. Depois das instâncias serem aglomeradas, são sujeitas a etiquetagem que auxilia a compreensão do significado de cada RS. As etiquetas variam em cada número de *cluster* atribuído, consoante o respetivo documento do par esteja colocado. O algoritmo 1 apresenta uma noção geral de como este passo se processa.

Neste processo, começa-se pela passagem das listas de contextos a um corpus através da *obterCorpus*. Esta função desempenha tal tarefa e mapeamentos no corpus a seguir. Remove as EN (consideradas exceções e que formem pares) e as palavras irrelevantes (na língua em que se encontram os dados) se requerido. Retira a pontuação, preservando os travessões entre palavras. Aplica stemização se exigida.

Usando o corpus, *pesarTermos* cria uma matriz documento-termo, com opções de controlo que são a função de *pesagem* e a tradução dos termos para letra minúscula, retirando antecipadamente documentos sem qualquer termo. A partir daqui, é computada a matriz de distâncias seguindo *medida_distancia*. Com esta estrutura e tendo o número exato de *clusters*, já se parte para um dos métodos de *clustering*: hierárquico ou k-Means (sendo este o escolhido por defeito). Assim toma-se conhecimento dos pares a que *cluster* pertencem por intermédio dos seus contextos.

Como ocorre em trabalhos anteriores, a atribuição de etiquetas a cada grupo de RS concretiza-se tendo em conta as frequências dos termos nos respetivos contextos[3, 5]. Neste trabalho, selecionam-se os *n_termos* mais frequentes (TF) do conjunto de contextos em cada grupo. Por fim, são associadas a cada par a chave do *cluster* a que pertence e a respetiva sequência de etiquetas. Aos pares sem *cluster*, atribuí-se 0 e nenhuma etiqueta. Na tabela 3.4, estão descritos os dados retornados.

3.5 Análise assintótica de *scanIteratively* e *scanEfficiently*

O método **scanIteratively** compara $\Theta(n)$ vezes se $only\_consecutive = FALSE$, sendo n o número de registos. Ou $O(n)$ vezes, se $only\_consecutive = TRUE$ e se

Algorithm 1 Definição de *clustering* para pares de EN com RS.

Entrada: Tuplos par, $pesagem$, $medida\_distancia$, $args = ()$, $args2 = (metodo = ward.D)$, $args3 = (algoritmo = Lloyd)$, $aglomerar = k - Means$, $n\_clusters$, n_termos e $remover\_stopwords$

Saída: Instâncias de par, *cluster* i e *etiquetas*

$corpus \leftarrow$ obterCorpus(par, $remover\_stopwords$)

$matriz\_pesos \leftarrow$ pesarTermos($corpus$, $pesagem$)

$matriz\_distancias \leftarrow$ dist($matriz\_pesos$, $medida\_distancia$, $args$)

Se $aglomerar == hierarquico$

  $arvore \leftarrow$ aglomerar hierarquicamente com $matriz\_distancias$ e $args2$

  $chaves \leftarrow$ cortar árvore em $n\_clusters$

Senão Se $aglomerar == k - means$

  $chaves \leftarrow$ obter $n\_clusters$ de *k-Means* com $matriz\_distancias$ e $args3$

Fim Se

Para i de 1 até $n\_clusters$ **Fazer**

  $etiquetas \leftarrow n\_termos$ mais frequentes por $matriz\_pesos$ e $chaves[i]$

  atribuir i e respetivas *etiquetas* a cada par por $chaves[i]$

Fim Para

Devolve par, i e *etiquetas*

Atributo	Descrição
entity1.name	Elemento de EN numa frase de texto processado.
entity2.name	Elemento distinto de EN na mesma frase de texto processado.
cluster.key	Chave do *cluster* ao qual o par pertence.
label.words	Etiquetas que identificam a RS do par.

Tabela 3.4: Pares de EN com as aglomerações de RS.

houver frases localizadas com apenas uma EN. Relativamente a $only\_consecutive = TRUE$, a comparação é feita $\Theta(nk)$ vezes, sendo k a quantidade máxima de EN presentes numa frase. O método pode chegar a $\Omega(n^2)$ quando todas as EN se encontram numa única frase.

Se $only\_consecutive = TRUE$, o método **scanEfficiently** é $\Theta(mk)$, tal que m é o número de frases em que aparecem EN e k a quantidade máxima destas numa frase. Na mesma condição, é $O(m)$ quando existe uma EN por frase ou é $O(k)$ quando todas estão na mesma frase. Por outro lado, se $only\_consecutive = FALSE$, é $\Theta(mk^2)$, ou $O(m)$ quando cada uma das frases tem só uma EN ou $O(k^2)$ quando há uma frase que tem todas as EN.

3.6 Sumário

Neste capítulo, apresentou-se uma abordagem com várias etapas até se aplicarem técnicas de *clustering*. Descreveu-se a solução ao nível do pré-processamento das EN, dos pares que estas formam e os seus contextos. Depois foi explicada a forma como se procede a aglomeração dos pares, sendo usados diversos recursos. Entretanto, detalharam-se informações acerca da implementação feita para este propósito. Nestas informações, destacam-se os conteúdos que quadros de dados podem ter durante o processamento. Ainda foram analisadas assintoticamente as duas funções definidas para a localização de EN a emparelhar. Como resultado, consegui-se software que permite aglomerar RS com diferentes configurações. No 4º capitulo, demonstra-se avaliação empírica que serve para encontrar as capacidades e as limitações da solução encontrada.

Capítulo 4

Avaliação Empírica

Neste capítulo, são demonstrados resultados obtidos de testes em 3 etapas diferentes do processo relativo à abordagem apresentada no capítulo 3. As etapas testadas são o reconhecimento de EN, o emparelhamento das mesmas e a aglomeração dos pares formados por estas com RS atribuídas. Na secção 4.1, é feita uma análise exploratória dos dados obtidos das duas primeiras etapas. A análise incide na distribuição das EN pelos diferentes parágrafos e frases dos documentos, que constituem o corpus usado para testar o algoritmo desenvolvido. Também incide nas frequências de pares de EN, que terão RS atribuídas. Pares esses que são criados consoante os critérios já conhecidos no capítulo 3.

A secção 4.2 indica a metodologia definida para avaliar a eficácia do algoritmo. Por fim, na secção 4.3, são apresentados resultados alcançados através da pontuação F1, relativos à aglomeração de pares de RS. Esses resultados dividem-se em duas partes principais: uma tem a ver com o contexto completo dos pares e a outra tem a ver com o contexto intermédio dos mesmos. Na primeira parte, são revelados também testes preliminares além dos que seguem a metodologia. A avaliação metodológica é desempenhada em duas fases, que serão detalhadas posteriormente na secção 4.2.

4.1 Análise Exploratória dos Dados

Para a abordagem implementada, usaram-se 227 textos não estruturados/anotados que são notícias publicadas no portal Sapo[1], passando por áreas desde informação internacional até economia, desporto ou sociedade. Foram disponibilizadas pelos Sapo Labs[29]. Não estão agrupadas por quaisquer critérios, tais como tempo, espaço ou temática. A publicação destas ocorreu no dia 31 do mês de Dezembro do ano de 2010. A coleção é usada para dois propósitos, principalmente para o último:

1. Encontrar falhas em várias funções que constituem a abordagem apresentada no capítulo 3.
2. Servir como input na elaboração dos testes de avaliação ao algoritmo de *clustering* proposto.

4.1.1 Reconhecimento de EN

Efetuou-se o reconhecimento automático das EN nos textos. Esta tarefa dura cerca de 1 minuto. Detetaram-se 6055 EN, sendo 2131 distintas. Contadas as EN, que estão no conjunto das frases onde cada uma é a primeira dum texto, tem-se 1096. Destas, 211 dizem respeito a *Lusa* ou *LUSA*. Tal EN refere uma fonte noticiosa[2] e aparece em cabeçalhos dos documentos. Em 205 casos, aparecem no primeiro parágrafo dum texto como pode ser visto abaixo:

> *Lisboa, 31 dez (Lusa) - O Ministério do Trabalho e da Solidariedade Social assegurou, na quinta-feira, que as chefias de quatro institutos da Segurança Social terão vencimentos mais baixos a partir de 2011.*

Quanto aos cabeçalhos doutros 21 ficheiros, *Lusa* aparece em parágrafos que seguem o primeiro dum texto, havendo avisos a anteceder. Assim está exposto um desses casos:

> **** Serviço vídeo disponível em www.lusa.pt ****
>
> *Viana do Castelo, 31 Dez (Lusa) - O Café do Repouso, em Chafé, Viana do Castelo, cumpre no sábado uma tradição de Ano Novo, que começou*

[1] http://www.sapo.pt/noticias/ - Acedido em 03/11/2016
[2] www.lusa.pt/ - Acedido em 03/11/2016

> *precisamente há 30 anos, disponibilizando comida e bebida de borla para todos, sejam ou não clientes da casa.*

Também a fonte aparece mencionada em rodapé nos textos, juntamente com a sinalização do final de artigo quase sempre desta forma:

> *Lusa/Fim.*

Cabeçalhos ou rodapés, tendo mais de uma EN reconhecida, poderão ter identificados pares para RS. Aliás, na demonstração do rodapé acima, está reconhecida a nomeação de *Fim*. Do mesmo modo aparece *FIM*. Estas duas situações sucedem-se 119 vezes sempre na primeira frase dalgum parágrafo. A hipótese de estes termos apenas servirem para assinalar o final dum artigo é evidente. Todavia, o facto de não haver outras palavras, para formar o contexto, facilita uma posterior remoção dos pares.

Quanto a EN no cabeçalho, podem ser úteis para a formação de pares. Nas amostras acima, são visíveis menções a localidades. Estas podem ser reconhecidas e emparelhadas como quaisquer outras. No entanto, as EN aí reconhecidas podem emparelhar com menos sentido porque não se inserem na estrutura duma frase.

Foram processadas 2488 frases, que se caracterizam por terem EN reconhecidas. O sumário das quantidades de EN, distribuídas por essas frases, é apresentado na tabela 4.1. Há uma ideia mais precisa dessa distribuição no gráfico da figura 4.1.

Minímo	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
1	1	2	2,43	3	17

Tabela 4.1: Números de EN reconhecidas por frase.

Existe um domínio de frases com 1, 2 ou 3 EN, pois constituem cerca de $78,42\%$ das frases processadas. Chegam a ser reconhecidas até 17 numa só frase. As frases que têm EN únicas são 1093, faltando 151 para formar metade e 1074 não pertencendo ao primeiro parágrafo dum texto. Das últimas, 95 não são a primeira frase dos respetivos

parágrafos. Se as EN únicas em frases forem excluídas no pré-processamento, haverá ganho na performance do algoritmo.

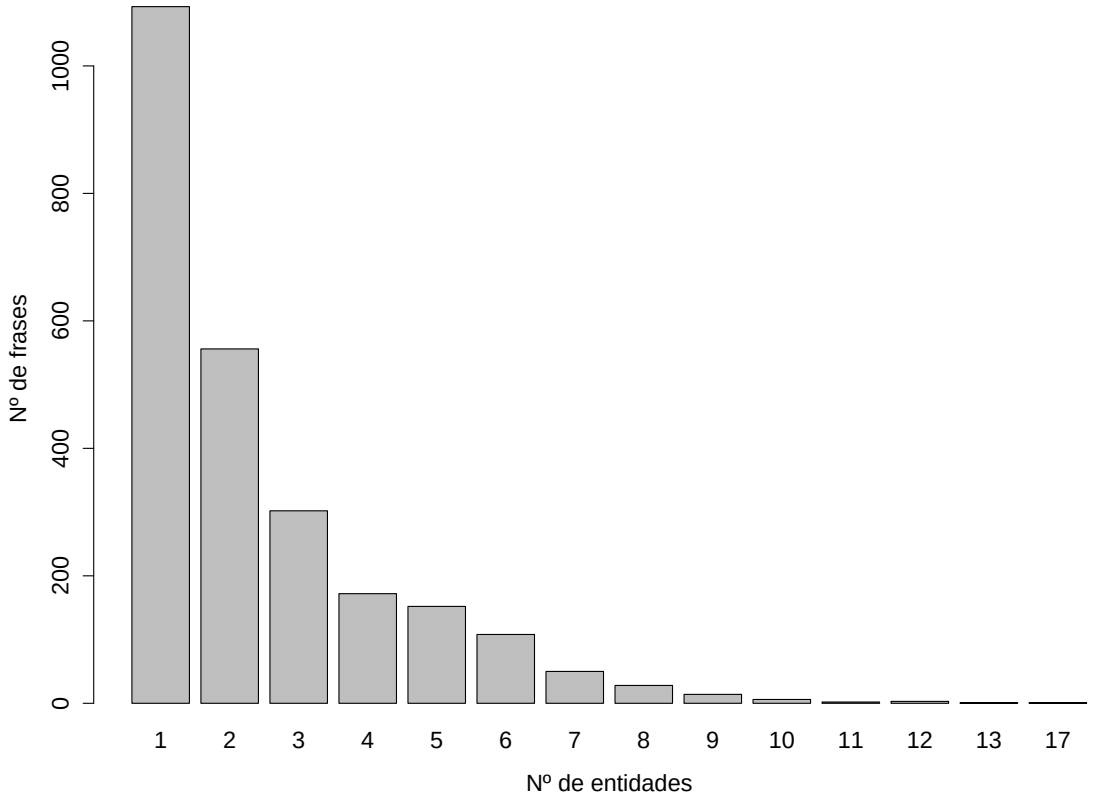


Figura 4.1: Quantidades de frases por número de EN nestas reconhecidas

Em relação a frases com mais do que EN únicas reconhecidas, são as que têm 2 EN que estão em maioria (relativa). O que se considera um bom ponto de partida para trabalhar, se forem usados os contextos por completo. Contudo, não deverão ser excluídas as que têm 3 ou mais EN, se se pretende obter resultados mais abrangentes em termos da diversidade de pares.

Consideraram-se *Lusa* e *Fim* desinteressantes. Removidos estes termos, relativos à fonte e à sinalização de fim dum artigo, passam a haver 5421 EN disponíveis. As EN removidas passam a ser denominadas como EN desinteressantes (END), o que significa que não possuem relevância semântica para o *clustering* quando incluídas no emparelhamento.

4.1.2 Emparelhamento de EN

Pares pela ordem de nomeação antes da remoção das END

Inicialmente, apenas se removeram 2 registos relativos a EN, devido aos seus valores do atributo específico não corresponderem às respetivas menções no texto. Depois deste pré-processamento, 6053 EN serviram para a formação de 3486 pares pela ordem de nomeação, cujos elementos dispõem-se consecutivamente na mesma frase.

Discute-se a distribuição dos pares formados, incluindo os das END, para ser demonstrada a diferença estabelecida entre as EN que são reconhecidas e as que são consideradas relevantes para *clustering*. Desta forma, realça-se a importância do pré-processamento em relação a EN extraídas.

O gráfico 4.2 transmite uma ideia das frequências de pares distintos. A tabela 4.2 mostra de forma evidente os pares mais frequentes desta seleção. Nota-se que o 1º e o 2º par se destacam, por várias dezenas de ocorrências, no que diz respeito à maioria. Alguns serão desnecessários para a aglomeração, como o caso do mais frequente, pois os respetivos contextos são insignificantes.

Par	**Frequência**
Lusa - Fim	117
Lisboa - Lusa	73
Porto - Lusa	14
Brasil - Itália	9
Cavaco Silva - BPN	9
Lusa - Governo	9

Tabela 4.2: Pares de EN mais frequentes com os respetivos elementos ordenados por menção (consecutiva), antes da remoção das END.

Combinam-se agora todos os pares possíveis que partilham uma frase. São 8574 no total. A figura 4.3 mostra a distribuição desses pares por frequência. Na tabela 4.3 são apresentados os pares mais frequentes nesta condição. Verifica-se o mesmo fenómeno

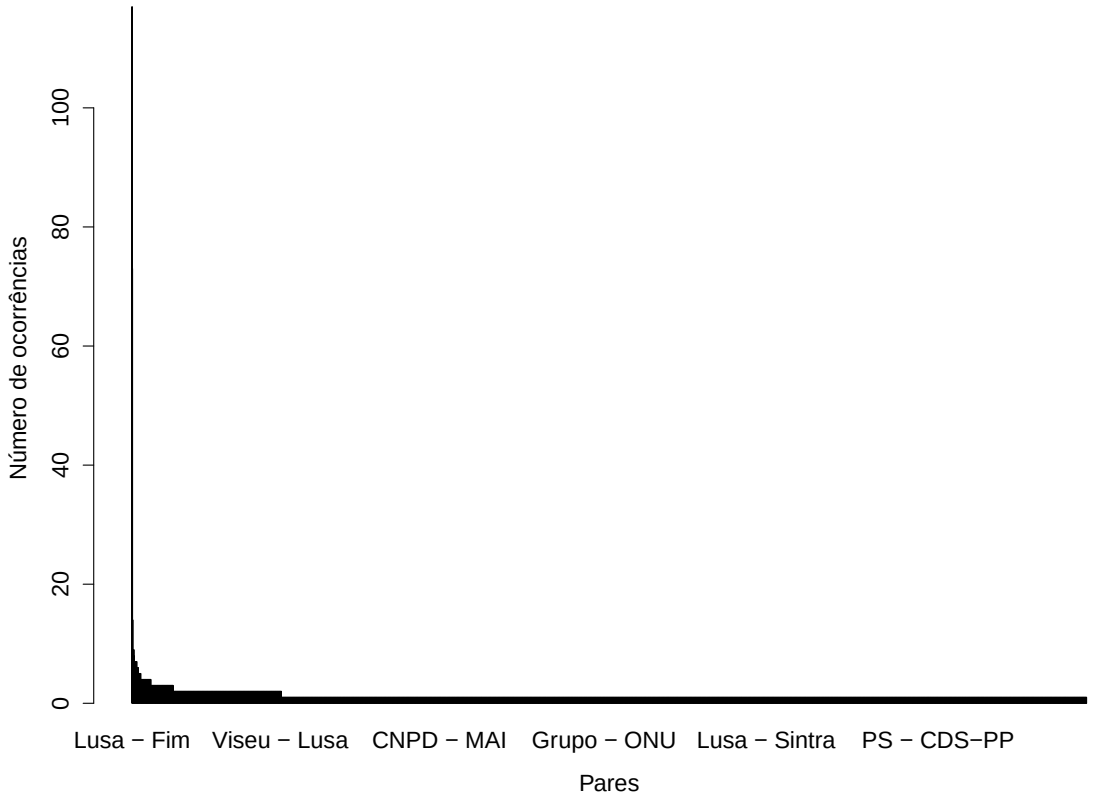


Figura 4.2: Distribuição das frequências dos pares de EN (apresentado exemplos) com os elementos ordenados por menção (consecutiva), antes da remoção das END.

da tabela 4.2, embora haja uma mudança em unidades no número de ocorrências tanto que do 3º ao 6º lugar estejam outros pares.

Pares pela ordem de nomeação após a remoção das END

A partir de 5421 que sobram da remoção das END, são obtidos 3071 pares pela ordem de nomeação, cujos elementos dispõem-se consecutivamente na mesma frase. Uma menor dispersão entre pares menos e mais frequentes é agora notada na figura 4.4. Agora na tabela 4.4 vê-se que os dois pares mais frequentes são outros e a diferença entre todos os mais frequentes é de 2 ocorrências. Existe portanto uma diminuição dessa dispersão.

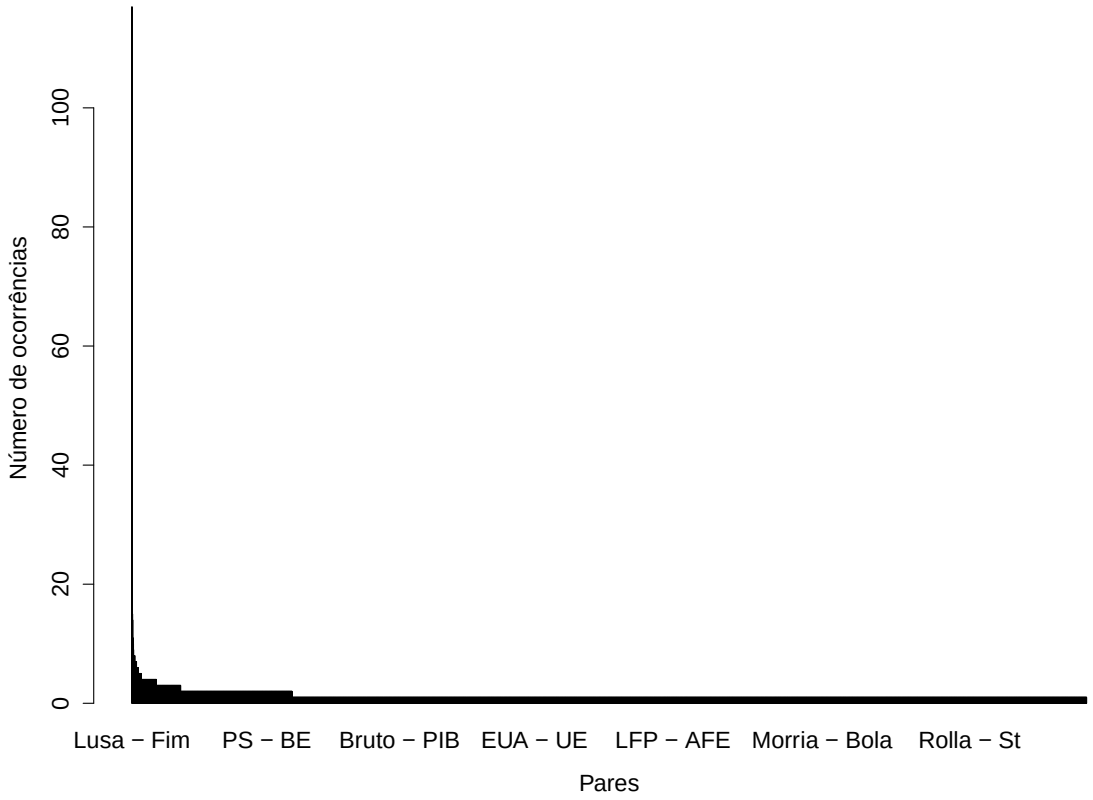


Figura 4.3: Distribuição das frequências dos pares de EN (apresentado exemplos) com os respetivos elementos ordenados por menção, antes da remoção das END.

Considerando inclusive os não consecutivos, obtém-se 7351 pares. A figura 4.5 demonstra a distribuição desses pares por número de ocorrências. Existe uma diferente disposição dalguns pares entre os mais frequentes, como é apresentada na tabela 4.5, onde as frequências aumentam algumas unidades.

Pares por ordem lexicográfica após a remoção das END

Por último, a análise é feita sobre os pares criados após a remoção das END, sem a ordem dos seus elementos nas frases ter influência. Portanto, considera-se o mesmo par tanto $a - b$ como $b - a$. As EN usadas para este emparelhamento são as mesmas 5421. O total dos pares determinados é 3071, tal como o dos pares formados nas mesmas condições com a exceção de se ter em conta a ordem lexicográfica dos seus elementos. O mesmo número deve-se ao facto de só a ordem dos elementos de certos pares ter sido trocada.

Par	**Frequência**
Lusa - Fim	117
Lisboa - Lusa	76
Lusa - Governo	21
Governo - Segurança Social	16
Porto - Lusa	15
Cavaco Silva - BPN	14

Tabela 4.3: Pares de EN mais frequentes com os elementos ordenados por menção, antes da remoção das END.

Par	**Frequência**
Brasil - Itália	9
Cavaco Silva - BPN	9
ministro dos Negócios - Estrangeiros	8
Battisti - Itália	7
Governo - Diário da República	7
Itália - Battisti	7

Tabela 4.4: Pares de EN mais frequentes com os elementos ordenados por menção (consecutiva), depois da remoção das END.

Por outro lado, já se nota uma distribuição diferente por frequência na figura 4.6, e na tabela 4.6 há exemplos dessa diferença. Pares anteriormente apresentados, com os seus elementos nas posições invertidas, têm agora a mesma frequência ou superior. Isto fez com que surgissem novos pares entre os mais frequentes.

Por exemplo, o par *Battisti* − *Itália* soma agora ao seu total a quantidade de pares *Itália* − *Battisti* como pode ser lembrado na tabela 4.4. Esta evidência demonstra a possibilidade de mais contextos poderem ser aglomerados, quando se considera a ordem lexicográfica, e assim o *clustering* prosseguir outro caminho.

Incluindo os não consecutivos, também são 7351 os pares determinados sob as mesmas condições, quando se tinha em conta a ordem das EN nas frases. A figura 4.7 mostra a distribuição das frequências destes pares. A tabela 4.7 demonstra mais um acréscimo de pares entre os mais frequentes e alterações nessas posições.

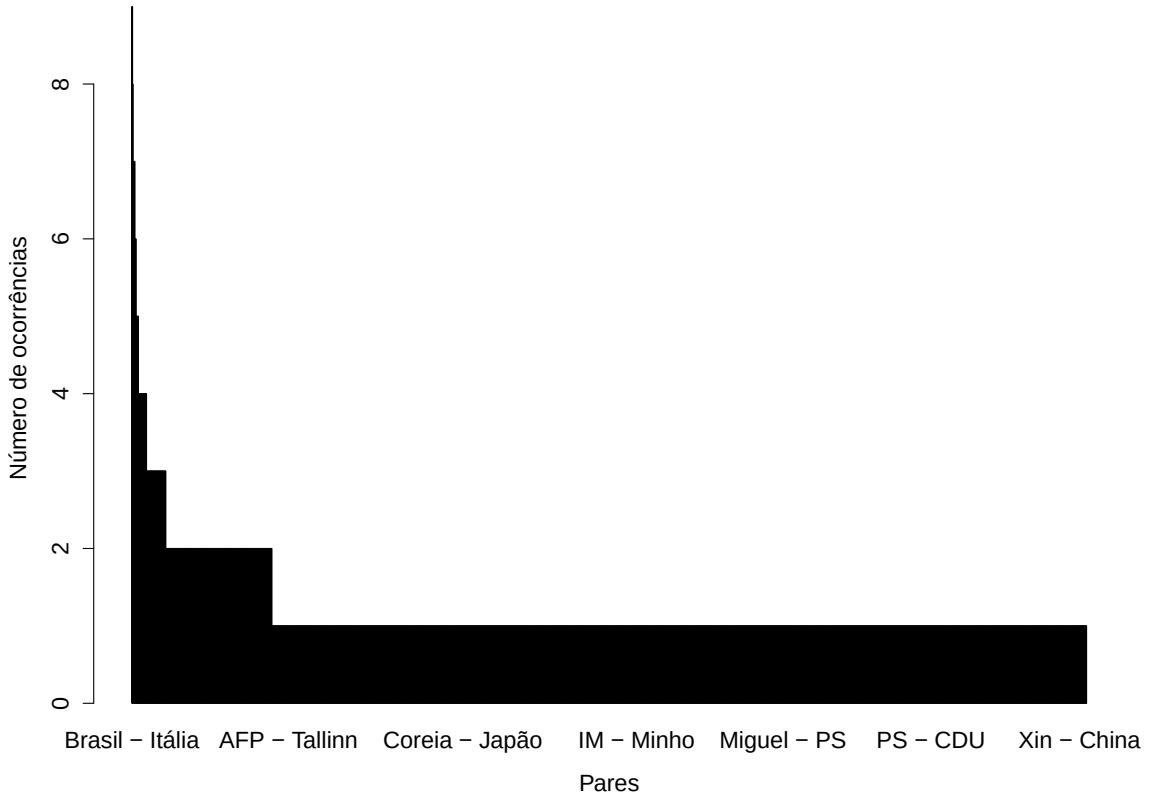


Figura 4.4: Distribuição das frequências dos pares de EN (apresentado exemplos) com os respetivos elementos ordenados por menção (consecutiva), depois da remoção das END.

As diversas condições no emparelhamento provocam diferentes conjuntos de pares. Sendo assim, tais condições poderão influenciar a qualidade da aglomeração de RS porque os contextos, que digam respeito a um determinado par, podem variar no seu número. Também esse factor tem influência, se houver opções de obtenção de contexto completo ou entre EN.

4.2 Definição da Avaliação

É definida uma metodologia de avaliação quantitativa na subsecção 4.2.3. Esta serve para considerar a distribuição dos pares de EN com RS entre si, tendo em conta várias quantidades de *clusters*. Antes da metodologia, serão definidos conceitos fundamentais para uma avaliação de *clustering* na subsecção 4.2.1. Também o formulário, necessário

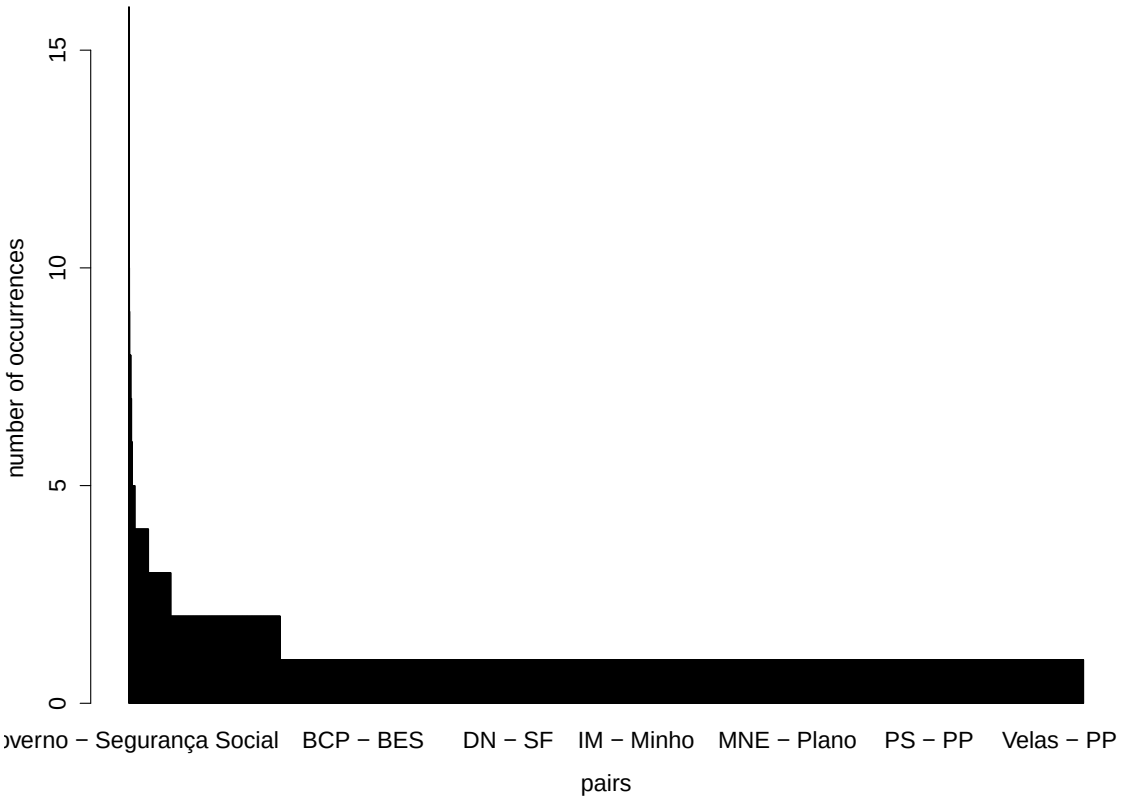


Figura 4.5: Distribuição das frequências dos pares de EN (apresentado exemplos) com os respetivos elementos ordenados por menção, depois da remoção das END.

à avaliação, é definido na subsecção 4.2.2.

4.2.1 Restrições sobre Pares

De modo a entender se um par de pares de EN deve partilhar um mesmo *cluster*, há que ter em consideração que esse par pode ser de um de dois tipos: $must - link$ e $cannot - link$. O primeiro é considerado quando os elementos desse par devem pertencer ao mesmo *cluster*. Já o segundo significa que os elementos desse par não devem fazer parte do mesmo *cluster*.

Os dois tipos de par são conceitos que serviram como constrangimento no emparelhamento, para aglomeração semi-supervisionada acompanhada de aprendizagem ativa, num trabalho apresentado[27] no capítulo 2. Contudo, nesta dissertação os tipos

Par	**Frequência**
Governo - Segurança Social	16
Cavaco Silva - BPN	14
Brasil - Itália	11
Diário da República - Segurança Social	10
Macau - China	9
Battisti - Brasil	8

Tabela 4.5: Pares de EN mais frequentes com os respetivos elementos ordenados por menção, depois da remoção das END.

Par	**Frequência**
Battisti - Itália	14
BPN - Cavaco Silva	14
secretário de Estado - Segurança Social	11
China - Macau	10
Brasil - Itália	9
Brasília - Olímpia	8

Tabela 4.6: Pares de EN mais frequentes com os respetivos elementos (consecutivos em menção) ordenados lexicograficamente, depois da remoção das END.

terão um significado mais específico. Não terão influência na distribuição dos pares de EN pelos *clusters*. Antes serão formados a partir de pares de EN que se sujeitarão a *clustering*, seguindo o propósito de apenas determinarem a qualidade do *clustering*.

Must-link: *Par cujos elementos* p_1 *(formados por* e_{left_1} *e* e_{right_1}*) e* p_2 *(formados por* e_{left_2} *e* e_{right_2}*) são pares de EN e que se considera encontrarem-se semanticamente ligados, sendo* $e_{left_1} \neq e_{left_2} \neq e_{right_2}$ *e* $e_{right_1} \neq e_{left_2} \neq e_{right_2}$.

Cannot-link: *Par cujos elementos* p_1 *e* p_2 *são pares de EN e que se considera encontrarem-se semanticamente desligados, sendo* $e_{left_1} \neq e_{left_2} \neq e_{right_2}$ *e* $e_{right_1} \neq e_{left_2} \neq e_{right_2}$.

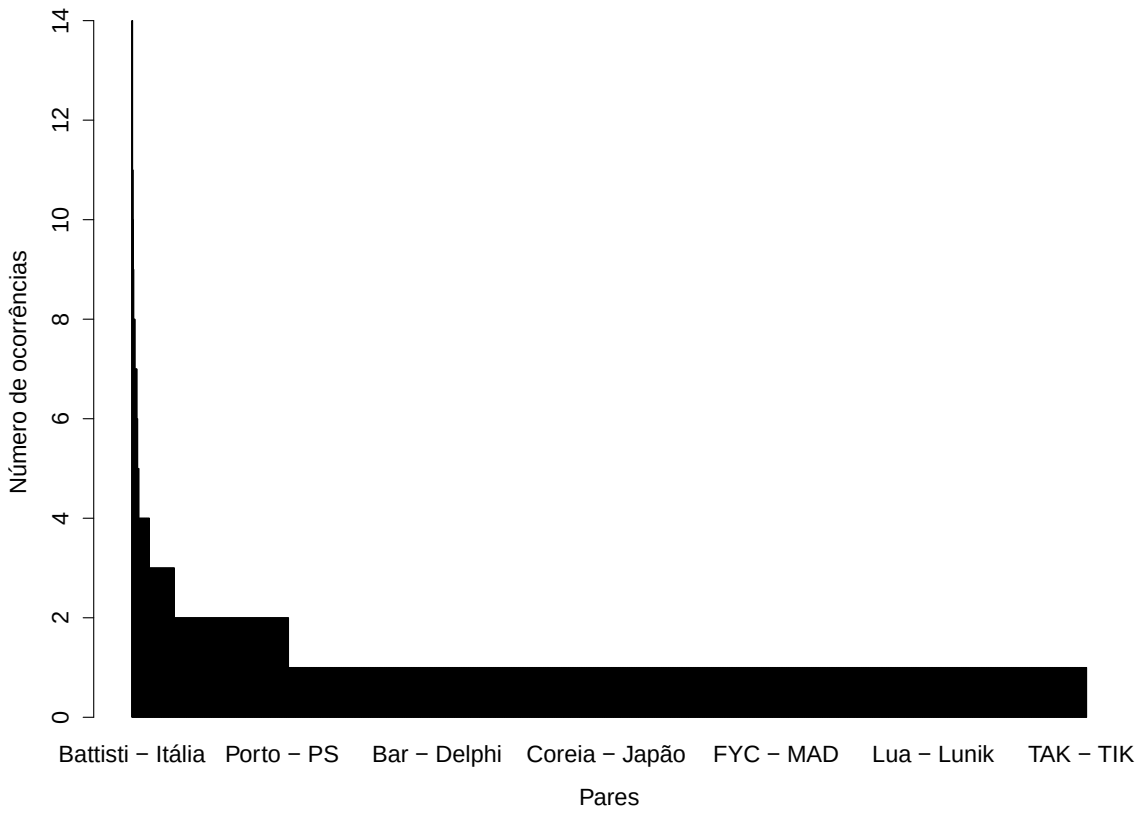


Figura 4.6: Distribuição das frequências dos pares de EN (apresentado exemplos) com os respetivos elementos (consecutivos em menção) ordenados lexicograficamente, depois da remoção das END.

Caso haja possibilidade de formar um $must - link$, através do par de EN $a - b$, com $c - d$ ou o seu inverso $(d - c)$, é escolhido o par que apresentar o antecedente e o consequente da ligação na mesma ordem em que se dispõem os elementos de $a - b$.

4.2.2 Avaliação de Aglomeração

A eficácia dos testes à solução proposta, para o agrupamento de pares com RS semelhantes, é quantificada por precisão e *recall*. Depois será calculada F1 combinando as duas medidas anteriores. Com esta medida, pretende-se entender até que ponto se obtém a melhor combinação entre precisão e *recall*. Antes de cada aglomeração, devem ser formados um conjunto de $must - link$ e outro de $cannot - link$ considerando:

- $\#TotalParesPrevistosNoMesmoCluster$ (PMC) como os pares que o método

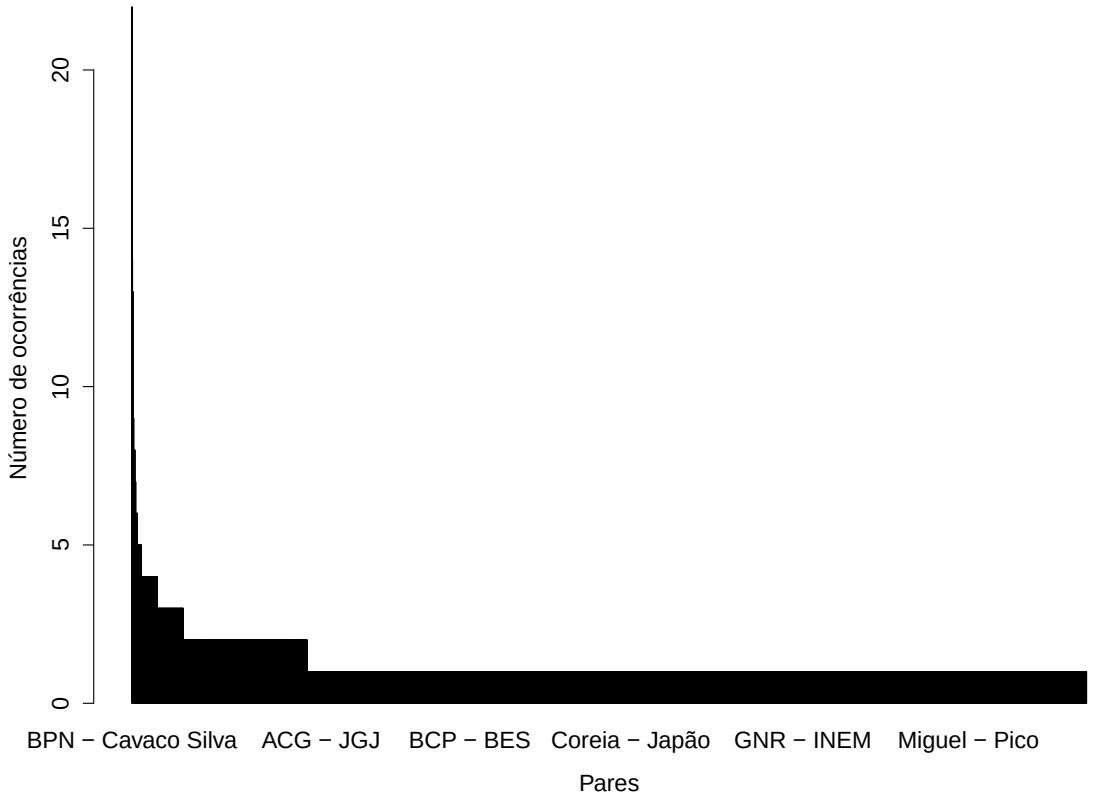


Figura 4.7: Distribuição das frequências dos pares de EN (apresentado exemplos) com os respetivos elementos ordenados lexicograficamente, depois da remoção das END.

indicará que pertencem ao mesmo *cluster*.

- $\#TotalParesAtualmenteNoMesmoCluster$ (AMC) como os pares determinados manualmente a estarem no mesmo *cluster*.
- $\#ParesCorretamentePrevistosNoMesmoCluster$ (CMC) como os que tiverem a previsão de *cluster* do método de acordo com a determinação manual.

Se são considerados X pares $cannot-link$ e M pares $must-link$, então XMC e MMC são X os $cannot-link$ indevidamente e M os $must-link$ devidamente previstos num mesmo *cluster*.

Por outro lado, XDC e MDC são X os $cannot-link$ devidamente e M os $must-link$ indevidamente previstos em diferentes *clusters*.

Par	**Frequência**
BPN - Cavaco Silva	22
Governo - Segurança Social	18
Brasil - Itália	15
Battisti - Itália	14
BPN - SLN	13
China - Macau	13

Tabela 4.7: Pares de EN mais frequentes com os respetivos elementos ordenados lexicograficamente, depois da remoção das END

Desta forma, tem-se então $X = XMC + XDC$ e $M = MMC + MDC = CMC + MDC$. Ou seja, $MMC = CMC$. Ainda há a reter que $PMC = XMC + MMC$ e $AMC = M$.

Portanto, as medidas para testar são formuladas da seguinte maneira:

$$Precision = CMC/PMC$$

$$Recall = CMC/AMC$$

No primeiro caso, obtém-se a percentagem de verdadeiros positivos em função de todos os positivos. No segundo caso, obtém-se a percentagem de verdadeiros positivos em função de todos os verdadeiros.

A medida final é calculada desta forma:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

4.2.3 Metodologia Aplicada

O método, a usar nos testes, que permanecerá inalterado é o k-Means para o processo de *clustering*, com 30 iterações permitidas no máximo. O método hierárquico foi preterido, ao invés do k-Means, pelo acréscimo da dificuldade em selecionar alturas onde árvores resultantes devem ser cortadas. De modo a haver maior hipótese de convergência em relação a 10 iterações, o número escolhido é 30.

Igualmente inalterados estarão outros parâmetros: a remoção das palavras comuns e o uso de stemização. Estes parâmetros justificam-se pela redução do número de termos. Assim evita-se que a matriz de distâncias fique tão pesada.

Outros 4 parâmetros, que combinam diferentes formas da execução do método, são a pesagem das frequências dos termos nos documentos, a medida de distância entre linhas da matriz, a quantidade de *clusters* e o algoritmo de aglomeração.

Diferentes combinações, de valores dos parâmetros, serão comparadas. Quando o contexto completo dos pares for usado, serão as 10 palavras mais frequentes do respetivo conjunto de contextos que etiquetam cada RS. Caso seja o contexto intermédio, serão apenas as 6 palavras mais frequentes porque é expetável que hajam menos palavras para etiquetagem.

Procura-se um equilíbrio no número de termos para etiquetagem. Por um lado, é importante haver etiquetas suficientes para a distinção/caracterização dos *clusters*. Pelo outro, deve-se evitar etiquetas que não sejam suficientemente representativas.

Definem-se c quantidades de *clusters* baseadas em percentagens do total de pares sujeitos ao agrupamento. Há uma quantidade intermédia usada apenas na 1ª execução enquanto as restantes usam-se na 2ª execução. As percentagens não são superiores a 50% para que haja pelo menos duas instâncias por grupo em média.

Esta separação serve para evitar resultados enviesados. A explicação destes resultados é que, antes da 2ª execução, são escolhidos mais $must - link$ e $cannot - link$ a partir dos *clusters* surgidos na 1ª execução. Perante isto, a avaliação relativa à quantidade intermédia teria uma inflação da pontuação devido à provável disposição repetida dos novos pares.

É permitida a escolha aleatória de centros iniciais, na prática do algoritmo k-Means, por se ter adotado uma estratégia não supervisionada. Por isso, os valores definitivos de F1 serão representados pela média aritmética simples dos valores obtidos nas tentativas válidas. Para se obter as estimativas de F1 em cada cenário, são comparadas as chaves de *cluster* entre os elementos de cada par $must - link$ ou $cannot - link$.

Uma tentativa é considerada válida, quando numa combinação de parâmetros não há 0 de F1. Esta condição é imposta devido a possíveis baixos valores da pontuação. Caso nenhuma tentativa resulte nalguma combinação, F1 é 0 para essa.

Portanto, haverão i tentativas por todas as combinações de parâmetros estabelecidos na execução inicial. Na execução final, serão desempenhadas j tentativas.

A avaliação processa-se com os seguintes passos:

1. Escolhem-se manualmente pares $must-link$ e $cannot-link$ iniciais. Os segundos deverão ser aproximadamente metade dos primeiros, de modo a dar-se relevo aos pares que coincidem através do *cluster*.
2. Executam-se aglomerações tendo em conta C_0 *clusters* e a medida de distância D_0, e calcula-se F1 dos testes resultantes com todas as combinações das pesagens $W^{1^{st}}$ e dos algoritmos A.
3. Escolhem-se manualmente $must - link/cannot - link$ adicionais a partir dos *clusters* resultantes, com o auxílio das etiquetas, relativos à melhor aglomeração em $W_0^{1^{st}}$, A_0 e i. Procura-se que o acréscimo não seja mais de metade dos iniciais, para que a etiquetagem tenha uma influência minoritária na avaliação.
4. Executa-se aglomerações com todas as combinações de C (exceto com C_0) e D, e calcula F1 para todos os cenários de $W^{2^{nd}}$ e A.

4.3 Testes

De modo a avaliar a implementação, são usados os mesmos textos como já tinha sido dito. Só se terá em consideração F1 como medida na avaliação metodológica. Nos testes realizados, os parâmetros terão os seguintes valores:

- Pesagens dos documentos na 1ª execução - $W^{1^{st}} = TF, TF/IDF, Binary, SMART$
- Pesagens dos documentos na 2ª execução - $W^{2^{nd}} =$*As duas pesagens que tiverem melhor média de F1 na 1ª execução. Em caso de igualdade, as que tiverem maiores F1 por um dos valores correspondentes a A. Em caso de nova igualdade, selecionar pesagens ao acaso.*

- Variações do k-Means - $A = Hartigan\ and\ Wong, Lloyd, Forgy, MacQueen$
- Números de *clusters* à percentagem - $C = 2.5\%, 5\%, 7.5\%, 10\%, 20\%, 30\%, 40\%$
- Medidas de distância - $D = euclidean, maximum, manhattan, canberra, binary, minkowski$

Ainda há a garantir que $C_0 = 10\%$, $D_0 = euclidean$, $i = 6$ e $j = 3$.

4.3.1 Contexto Completo

Nesta fase, todos os testes ao *clustering* de RS serão executados com o aproveitamento dos contextos completos ($what\_context = all$). Começa-se pelos pares únicos em cada frase, que no total são 380. As EN dos seus elementos dispõem-se pela ordem lexicográfica, após a remoção das END (*Lusa* e *Fim*).

4.3.1.1 Testes Preliminares

Foram feitos dois testes de avaliação introdutórios. Serviram para entender o comportamento da abordagem ao nível da execução do algoritmo proposto. Também estes testes serviram para apurar a metodologia definida. Por fim, a utilidade dos mesmos chega ao auxílio na escolha dos $must - link$ e $cannot - link$.

Para a realização do primeiro teste preliminar, serão usados $W = TD/IDF$, o algoritmo de k-Means ($A = Lloyd$) com permissão de 10 iterações no máximo, $D = euclidean$, $C = 10\%$ e remoção de *stopwords*. De resto, os parâmetros são os que se encontram por defeito embora sejam escolhidas também as 10 palavras mais frequentes de cada *cluster* para identificar as respetivas RS.

Identificaram-se 20 $must - link$ anotados humanamente. Para cada um dos pares, foi anotada a razão de cada associação e palavras-chave. A tabela 4.8 apresenta os seus pares.

Depois identificaram-se 10 $cannot - link$ da mesma forma. Incluíram-se as razões para as RS distintas de ambos os elementos de cada par. A tabela 4.9 apresenta esses pares.

1º elemento		2º elemento	
1º elemento	2º elemento	1º elemento	2º elemento
Grécia	Portugal	África do Sul	Moçambique
Evo Morales	Presidente da Bolívia	Hugo Chávez	presidente da Venezuela
Cristiano Ronaldo	Luís Figo	Di Maria	Nico Gaitán
Argentina	Buenos Aires	Cairo	Egipto
Matias Fernández	Paulo Sérgio	André Villas-Boas	Falcao
Sporting de Braga	Vitória de Setúbal	Benfica	Marítimo
Facebook	Twitter	Portugal Telecom	Vivo
Bertie Ahern	Irlanda do Norte	Brasil	Malam Bacai Sanhá
Berlusconi	Itália	Cameron	Reino Unido
Abdoulaye Wade	África	Europa	Sarkozy
BPP	João Rendeiro	BCI	Paulo Manhique
ACNUR	Libéria	Conselho dos Direitos Humanos da ONU	Costa do Marfim
Dallas Mavericks	Dirk Nowitziki	Benfica	Francisco Jordão
Barcelona-Dacar	Espanha	Portugal	Rali Lisboa-Dacar
Filipinas	Manila	China	Pequim
Esp	María José Pueyo	JESSICA AUGUSTO	POR
Alice Timbilili	Quénia	Eriteira	Zarsenay Tadese
Ayala	Racing Club	Adrian	Wigan
Croácia	Hungria	China	Japão
Bertie Ahern	Fianna Fail	Nicolau dos Santos Lobato	presidente da Fretilin

Tabela 4.8: Primeira formação de pares $must-link$ para avaliar *clustering* de pares únicos em frases.

1º elemento		2º elemento	
1º elemento	2º elemento	1º elemento	2º elemento
Açores	Madeira	África do Sul	Madjone-djone
Battisti	Itália	BPN	BPP
Altri	Inapa	Abbas	Dilma Rousseff
China	Macau	Bowler	Jerome Pelichet
Chui Sai On	Executivo de Macau	Dragão	FC Porto
Armando Guebuza	Moçambique	Anatoly Karpov	Gari Kasparov
Alassane Ouattara	Gbagbo	Grupo Leoni	Guimarães
China	Estados Unidos	diretor do GAVE	Hélder Diniz de Sousa
Brasil	Celso Amorim	Induta	Justiça Militar
África do Sul	Itamaraty	Mikhail Khodorkovski	Platon Lebedev

Tabela 4.9: Primeira formação de pares $cannot-link$ para avaliar *clustering* de pares únicos em frases.

A avaliação determinou 100% de precisão e 5% de *recall*. O que resulta em $9,52\%$ de F1. Contudo, um $must-link$ teve os seus elementos sem *cluster* atribuído. O tempo de execução é cerca de 5 minutos.

No segundo teste prévio, a avaliação é feita com mais $must-link$ e $cannot-link$. São 5 adicionados aos primeiros e são 12 $cannot-link$ que substituem os anteriores. No 2º conjunto destes pares, o critério de emparelhamento é idêntico ao que deve ser efetuado na metodologia após a 1ª execução. Ou seja, seguiu-se o resultado obtido num *clustering*, usando stemização e alguns parâmetros com valores alterados, enquanto outros mantiveram os seus valores. As anotações humanas que justificam ambos os

tipos de pares permaneceram.

No entanto, foram retirados dos pares, anteriormente sujeitos a *clustering*, 23 com pelo menos uma das EN a ser inválidas e 30 que não tinham sido atribuídos a qualquer *cluster*. Nas tabelas 4.10 e 4.11, são apresentados os novos pares.

1º elemento		2º elemento	
1º elemento	2º elemento	1º elemento	2º elemento
Maratona	Sara Moreira	Ercília Machado	SC Braga
Jogos Olímpicos de Atenas	Sérgio Paulinho	Pequim	Telma Monteiro
Bielorrússia	Lukachenko	Bertie Ahern	primeiro-ministro da Irlanda
Bertie Ahern	Dublin	David Cameron	Londres
António Leitão	Carlos Lopes	Carlos Mozer	Rogério Gonçalves

Tabela 4.10: Segunda formação de pares $must-link$ para avaliar *clustering* de pares únicos em frases.

1º elemento		2º elemento	
1º elemento	2º elemento	1º elemento	2º elemento
Anna Bligh	Estado de Queensland	Chefe de Estado	Manuel Alegre
Bloco de Esquerda Porto	João Teixeira Lopes	Azeddine Araba	Séfit
Constituição da Rússia	Kremlin	Pamir	Tajiquistão
Estádio Cidade de Coimbra	Liga	Diário da República	SIC
Associação Nacional de Defesa dos Clientes do BPN	Lisboa	Governo Regional dos Açores	Ponta Delgada
Fase Nova	Lua	Kennedy	Presidência dos EUA
Go Bulling	Luís Gonçalves	Amadora	Manuel Damião
Arnoia	Malaqueijo	La Familia	Nazario Moreno
Avenida dos Aliados	Manuel Clemente	Luís Gonçalves	PSI
Ensitel	Maria João Nogueira	Filipe Garcia	Jerónimo Martins
Caron Butler	Mavericks	II Guerra Mundial	Plano Marshall
Graeme McDowell	Mike Catt	ONU	Teerão

Tabela 4.11: Segunda formação de pares $cannot-link$ para avaliar *clustering* de pares únicos em frases.

A nova aglomeração determinou 60% de precisão e 6% de *recall*, diminuindo 40% e aumentando 1% respetivamente em comparação com a aglomeração anterior. Assim verificou-se um aumento ligeiro do F1 para $10,91$. A execução demorou cerca de 5 minutos, tal como no teste preliminar anterior. Coloca-se a hipótese do *trade-off* estar relacionado com os diferentes valores de vários parâmetros, com o pré-processamento e/ou com o critério de seleção dos pares $must-link$ e $cannot-link$.

4.3.1.2 Testes Metodológicos

Para se evitar resultados enviesados, decidiu-se usar o conjunto dos pares do 2º teste prévio com menos dois pares (considerados $must - link$) que não tinham *cluster* atribuído. Portanto, pares destes e outros, cujos elementos reconhecidos não sejam entendidos como EN reais, ficam de fora. O segundo caso significa palavras no início de frases que representam verbos/adjetivos ou separação errada de EN. A tabela 4.12 demonstra tais pares.

1º elemento	2º elemento
Argentina	Cerca
Cerca	Costa do Marfim
CEDEAO	Deste
diretor de Direção	Exceção
Assembleia da República	Fazendo
Feito	Gabinete Médico-Legal de Leiria
Conforme	IDT
Discuti	Korir
Estrangeiros	ministro dos Negócios
Deste	MMC
Hernâni Cidade	Morre
Manuel Rivera	Morre
Interior	Morria Liese Prokop

Tabela 4.12: Pares de EN, dos quais pelo menos um dos seus elementos foi considerado inválido, e portanto não serão usados para avaliar *clustering* de pares únicos em frases.

Retirados pares falsos ou com contexto insignificante, sobrando 325, é efetuada a 1ª execução com base na metodologia. Os resultados obtidos, com 32 *clusters* requeridos, são demonstrados na tabela 4.13. A pesagem TF prova ter a melhor pontuação média por pesagem, embora não se verifique isso com todos os algoritmos de k-Means. Contudo, foram obtidos os melhores resultados usando o algoritmo de Lloyd com a pesagem Smart.

Foi praticada a 2ª execução. A figura 4.8 apresenta resultados para os 4 algoritmos com a pesagem Smart. A maioria dos algoritmos demonstra que, com cerca de 10 *clusters*, os resultados são melhores (como é claro nas figuras 4.8b, 4.8c e 4.8d).

Nota-se um domínio da medida de distância máxima quando são requeridas as 3 quantidades menores de *clusters*. Com esta medida e com o algoritmo de Forgy

	Hartigan-Wong	**Lloyd**	**Forgy**	**MacQueen**	**média**
TF	7,93	7,20	7,90	6,02	7,26
TF/IDF	4,10	4,05	5,95	4,00	4,53
Binary	7,90	5,62	5,60	7,10	6,56
SMART	7,30	9,20	5,32	5,97	6,95

Tabela 4.13: Resultados de F1 por algoritmo/pesagem na 1ª execução da metodologia de avaliação, para pares únicos em frases com *what_context = all*.

(presente na figura 4.8c), o valor de F1 aproxima-se dos 50% a partir da menor quantidade requerida de *clusters*.

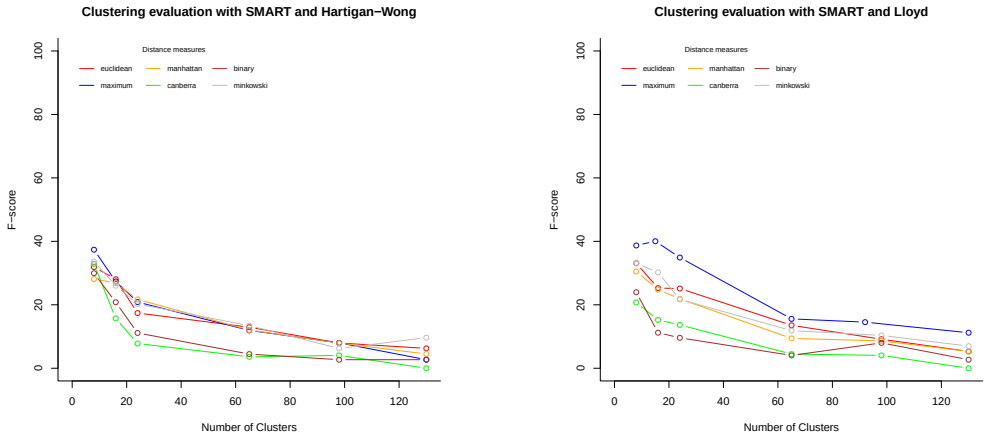
Na figura 4.9 são vistos os resultados baseados em TF. Confirmam-se algumas tendências registadas na pesagem anterior, nomeadamente, o domínio da distância máxima e o algoritmo de Forgy a aproximar-se mais dos 50% de F1. Desta vez, essa aproximação deveu-se à 2ª menor quantidade requerida de *clusters* (entre 15 e 20). É expetável que isso se deva a uma mudança discrepante de precisão e/ou *recall* em relação à diminuição de *clusters* requeridos.

4.3.2 Contexto Intermédio

Desta vez, com *what_context = between*, os elementos que formam os pares de EN encontram-se pela ordem que aparecem nas respetivas frases. A avaliação continua a seguir a metodologia. Apesar de haver menos palavras em contextos, mantém-se *stopwords* de fora por causa da possibilidade de aparecerem conjunções ou disjunções que enviesam a aglomeração de RS.

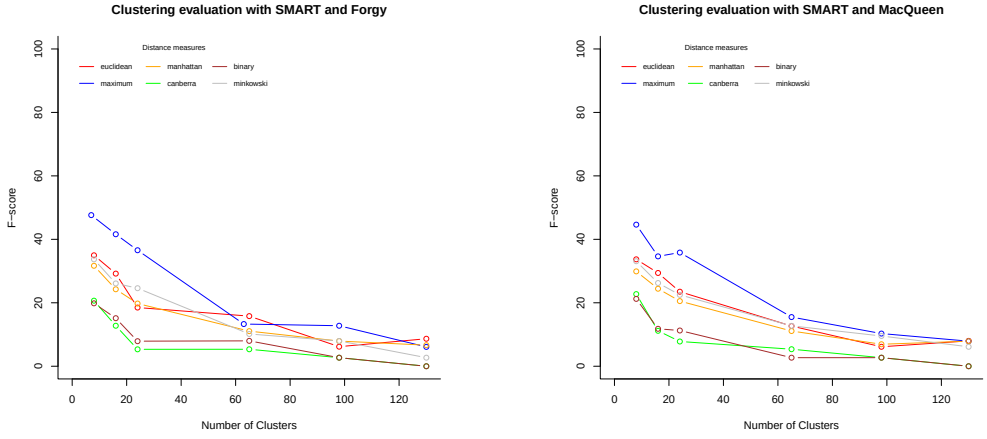
4.3.2.1 Testes Metodológicos

Tal como com o contexto completo, removem-se registos relativos às END e depois retiram-se pares sem *cluster* atribuído ou que têm elementos reconhecidos que não são entendidos como EN reais. Sobram 329 pares para testar. São mais 4 do que no conjunto dos pares, com *what_context = all*. Isto deve-se ao facto de agora haver alguns pares com as mesmas EN dispostas em ambas as ordens.



(a) Hartigan and Wong

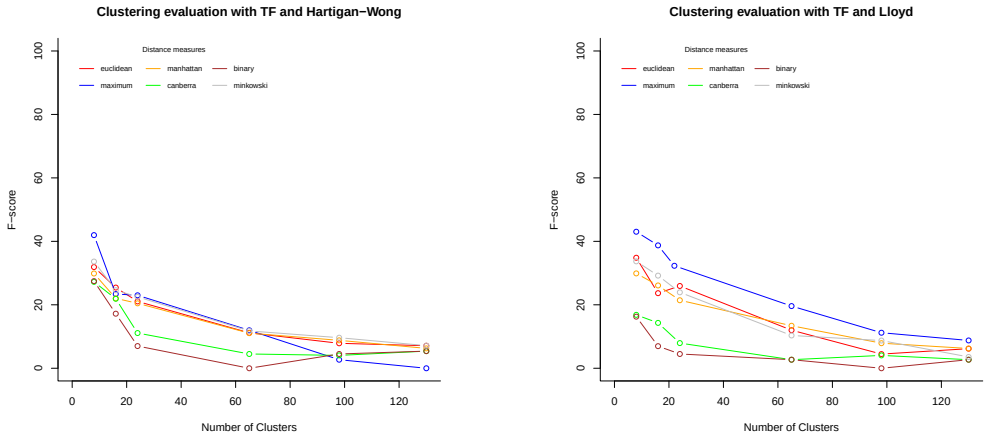
(b) Lloyd



(c) Forgy

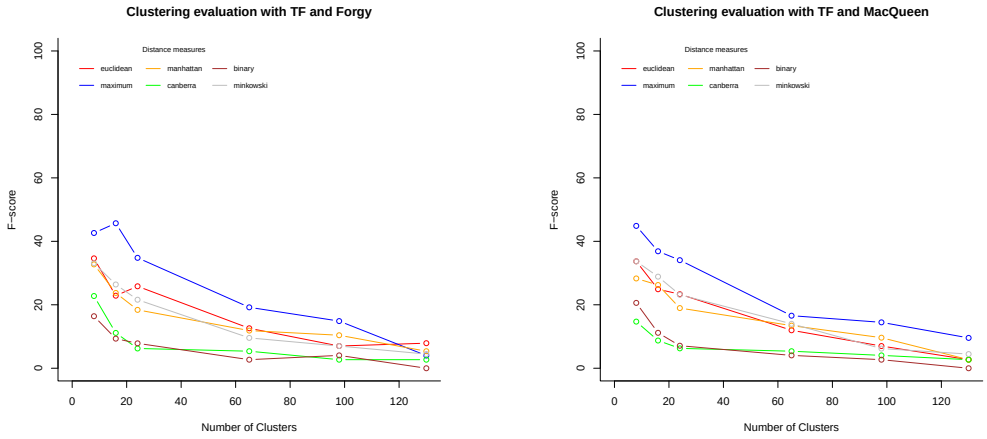
(d) MacQueen

Figura 4.8: Resultados de F1 em 6 quantidades requeridas de *clusters*, por algoritmo de k-Means, na 2ª execução da metodologia de avaliação, para pares únicos em frases com $what\_context = all$ e $W^{2^{nd}} = SMART$. Legenda mais percetível no anexo B.



(a) Hartigan and Wong

(b) Lloyd



(c) Forgy

(d) MacQueen

Figura 4.9: Resultados de F1 em 6 quantidades requeridas de *clusters*, por algoritmo de k-Means, na 2ª execução da metodologia de avaliação, para pares únicos em frases com $what\_context = all$ e $W^{2^{nd}} = TF$. Legenda mais percetível no anexo B.

Efetua-se a 1ª execução, com 33 *clusters* requeridos, onde os resultados obtidos são apresentados na tabela 4.14. A pesagem Smart prova ter a melhor pontuação média. No entanto, as melhores pontuações, com o mesmo valor, foram o algoritmo de Lloyd usando a pesagem TF/IDF e o algoritmo de Forgy usando a pesagem Smart.

É realçada também a diferença entre a maior e a menor pontuação média por pesagem (Smart e binária) que é de aproximadamente apenas $0,9$. Face a estes resultados, a melhor abordagem para aglomeração com 33 *clusters* é a que utiliza Smart e Forgy.

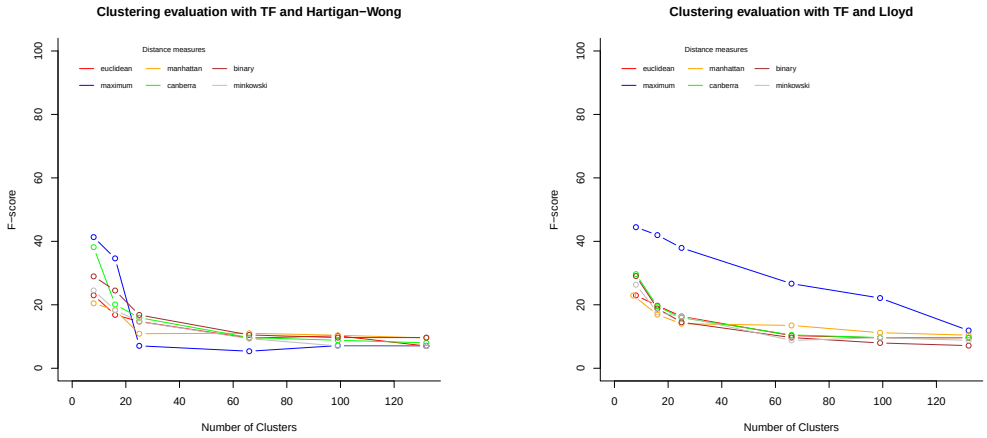
	Hartigan-Wong	**Lloyd**	**Forgy**	**MacQueen**	**média**
TF	9,27	10,53	9,27	9,27	9,58
TF/IDF	8,00	11,13	8,00	9,87	9,25
Binary	8,63	9,90	8,00	8,63	8,79
SMART	9,90	9,27	11,13	8,63	9,73

Tabela 4.14: Resultados de F1 por algoritmo/pesagem na 1ª execução da metodologia de avaliação, para pares únicos em frases com $what\_context = between$.

Na figura 4.10, relativa à 2ª execução, está destacada a medida de distância máxima, nos resultados baseados na pesagem TF, quanto menor é a quantidade de *clusters* requeridos. As figuras 4.10b, 4.10c e 4.10d demonstram isso mesmo, chegando a ser alcançada F1 entre 40% e 50%. Também a figura 4.10a prova que F1 da distância máxima chega a um valor idêntico, embora com maiores *clusters* requeridos fique com pontuação abaixo doutras medidas.

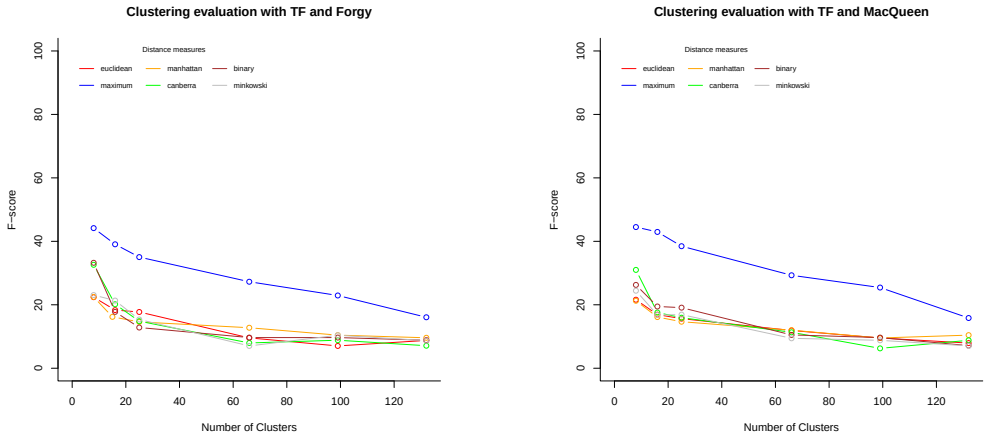
Na figura 4.11, relativa à 2ª execução, destaca-se a medida de distância máxima, nos resultados baseados na pesagem Smart, quanto menor for a quantidade de *clusters* requeridos. Ocorre igualmente para os mesmos algoritmos, demonstram as figuras 4.11b, 4.11c e 4.11d isso mesmo. Chega também a ser alcançada F1 entre 40% e 50%.

A figura 4.11a prova, o que se verificou na pesagem anterior. A distância máxima chega a um valor de F1 idêntico ao doutras medidas de distância. Com as 3 maiores quantidades requeridas de *clusters*, a mesma medida de distância chega a ter pontuação abaixo doutras medidas.



(a) Hartigan and Wong

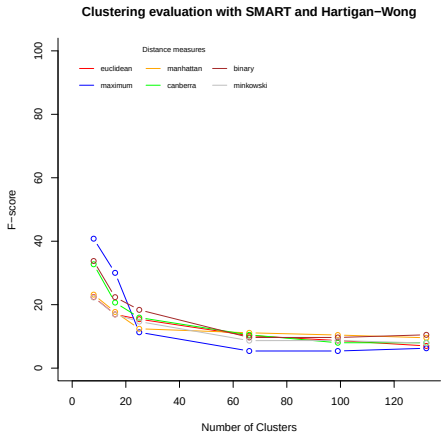
(b) Lloyd



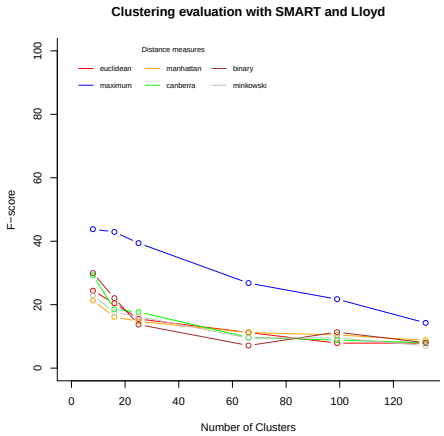
(c) Forgy

(d) MacQueen

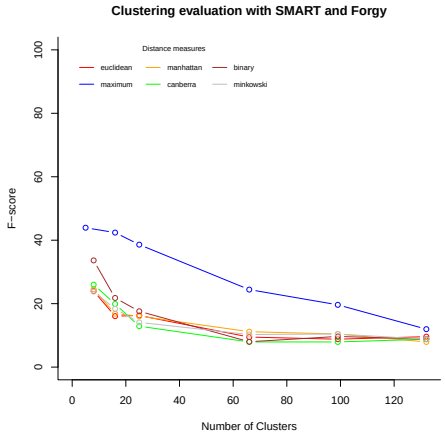
Figura 4.10: Resultados de F1 em 6 quantidades requeridas de *clusters*, por algoritmo de k-Means, na 2ª execução da metodologia de avaliação, para pares únicos em frases com $what\_context = between$ e $W^{2^{nd}} = TF$. Legenda mais percetível no anexo B.



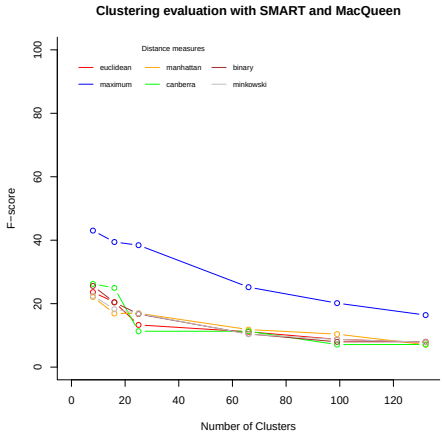
(a) Hartigan and Wong



(b) Lloyd



(c) Forgy



(d) MacQueen

Figura 4.11: Resultados de F1 em 6 quantidades requeridas de *clusters*, por algoritmo de k-Means, na 2ª execução da metodologia de avaliação, para pares únicos em frases com $what\_context = between$ e $W^{2nd} = SMART$. Legenda mais percetível no anexo B.

4.4 Sumário

A seleção de EN, consoante a sua frequência por frase, é importante para a definição de trabalho a desenvolver. Uma das razões está relacionada com os casos onde há mais do que um par de EN numa mesma frase e a extração do contexto seja completa. Deste modo e usando um saco-de-palavras, todos os pares duma frase teriam os mesmos termos, o que poderia levar a que se considerassem semelhantes quando isso nem sempre ocorre. Um exemplo onde a semelhança nas RS entre pares da mesma frase não acontece é o seguinte, tendo em conta que *Laranjas*, *Amarelos* e *Campeonato Universal* são EN:

> *Os* ***Laranjas*** *venceram os* ***Amarelos*** *que se estrearam no* ***Campeonato Universal****.*

Outra razão da importância de selecionar EN, pela frequência por frase, está no acréscimo expetável de possíveis pares das mesmas. Esta mudança provoca aumento do tempo e do espaço para processamento. O uso de milhares de pares, como por exemplo 2000, implicou aglomerações, idênticas às dos testes preliminares, com duração superior a 25 minutos. Num cenário em que estes milhares de pares seriam testados com várias combinações de parâmetros, como se sucede através da metodologia de avaliação proposta, seria atingido um tempo de execução inaceitável.

Por outro lado, com centenas de pares torna-se mais difícil o emparelhamento manual de $must - link$. Isto deve-se à menor probabilidade de se encontrar uma quantidade representativa destes tipos de pares. Sendo assim, torna-se mais difícil concretizar a avaliação eficaz e eficientemente.

Com foco nos testes metodológicos, é evidente que com menores quantidades requeridas de *clusters* são obtidos melhores resultados. Este cenário ocorre independentemente dos valores doutros parâmetros. No entanto, nota-se superioridade de F1 nalguns resultados que usam contextos completos em comparação com os que usam contextos intermédios.

A partir destes e outros factos analisados nos conjuntos de gráficos, pode-se dizer que o algoritmo de Forgy é o melhor para aglomerar RS. Mesmo que o contexto

intermédio por algum motivo tenha de ser usado, Forgy como algoritmo de k-Means atinge melhores resultados nalguns cenários.

Capítulo 5

Conclusão

Nesta tese, abordou-se o problema de agrupar automaticamente RS entre EN em textos. Este é um problema de grande relevância para a análise e a compreensão de textos feita por computador e que tem inúmeras aplicações. Utilizaram-se técnicas de *clustering*, numa abordagem que começa pelo reconhecimento das EN e prossegue com o emparelhamento das mesmas lexicograficamente ou por ordem em que aparecem nas frases.

A abordagem permite que a fase de aglomeração se suceda de diversas formas. Existe a possibilidade de mapear o corpus, que contém ou os contextos intermédios ou os contextos completos de pares, com stemização e/ou remoção de palavras. De seguida, a pesagem das frequências dos termos e as medições de distâncias entres estes são executadas conforme alguns parâmetros se encontram. Outros parâmetros permitem definir o algoritmo de *clustering* usado, que tanto pode ser hierárquico como de k-Means. A quantidade de etiquetas que caracteriza cada *cluster* criado é também parametrizada.

Estudaram-se os posicionamentos das EN em textos[29]. Isto serviu para determinar o uso apenas de pares únicos em frases, nos testes realizados ao algoritmo proposto. Depois de pré-processamento relativo a EN e aos pares que formam, foram obtidos resultados de testes com diversas configurações. Depois dos preliminares, fizeram-se novos testes que seguiram uma avaliação metodológica para ambas as extrações de contexto. Nestes testes compararam-se diferentes medidas de distância, pesagens e quantidades de *clusters*. Além disso, usado o k-Means, sendo comparados 4 tipos.

A avaliação metodológica que também foi apresentada, e usa pares de constrangimento $must - link$ e $cannot - link$, serviu para tirar ilações. Percebeu-se que há medidas de distância e pesagens que superam outras. Notou-se um aumento de F1 à medida que a quantidade requerida de *clusters* diminui. E o algoritmo de Forgy demonstrou melhor pontuação máxima em alguns cenários.

O trabalho de dissertação originou uma solução de software[30]. Esta pré-processa informação sobre EN, os pares que formam e seus contextos. Sobretudo permite explorar diversos caminhos que influenciam no final a aglomeração de pares com base nas suas RS. A solução escrita em R[16] usa **PAMPO**[1] para reconhecer as EN e lida com várias técnicas de *text mining*[28] e de *clustering*. É disponibilizada uma componente do software que serve para fazer a avaliação, quando inseridas *strings* que descrevem pares $must-link/cannot-link$ escolhidos manualmente. São disponibilizados também dados gerados a partir de diferentes etapas do processamento. Seguindo a abordagem apresentada em tese, consegue-se consultar vários relacionamentos a partir duma entidade sugerida.

5.1 Limitações e Trabalho Futuro

A solução utilizada para o reconhecimento de EN tem em conta a sua desambiguação. Contudo, há casos registados que dizem respeito a conceitos genéricos que a desambiguação não permite especificar. Por exemplo, a EN *Governo* não permite por si só saber de qual país é. Acontece o mesmo com *Presidente*, pois não se sabe que organismo é presidido. Uma ideia para solucionar esta questão passa por se explorar cabeçalhos dos textos ou os próprios contextos das EN de modo a especificarem-se conceitos. O que se pretende é que isso contribua para a separação dos contextos agregados que por sua vez evita a deturpação na obtenção de RS. No entanto, uma melhor desambiguação não evita a agregação de contextos que evidenciam diferentes RS entre o mesmo par de EN. Para esta situação, seria útil que houvesse uma distinção de contextos com base na semelhança entre os respetivos termos. Essa distinção poderia ser resolvida com um limiar de similaridade. Ou ser resolvida com uma solução mais sofisticada como Wordnet[23].

Ainda relativamente a melhorias na agregação de contextos, há a indicar que seria interessante testar este processo integrado no emparelhamento. Significa que, à medida

que os pares iam sendo criados, juntava-se cada novo contexto a outros pertencentes a pares já existentes. Por outro lado, a possibilidade de eliminar cabeçalhos e rodapés, que estejam juntos com primeiras e últimas frases de textos respetivamente, melhoraria não só a capacidade de agregação como a de emparelhamento.

Outra dificuldade notada é o conjunto dos textos testados serem relativos a publicações de apenas um dia e não estarem agrupados por domínio. Na primeira situação, a existência de mais variedade de pares ajudaria possivelmente na definição de $must-link/cannot-link$ e na qualidade das aglomerações. E na segunda situação, não só o que é indicado na primeira como a eventual diminuição de ruído, por contextos de diferentes domínios estarem separados, seria um cenário real.

A não etiquetagem das palavras de contexto dificulta a definição de RS. De certa forma, o processo de POST auxiliaria na discriminação de termos com base na sua classe gramatical. Além disso, palavras menos próprias como adjetivos, serviram indevidamente para a caracterização de pares, após as tarefas de aglomeração. Isto dificultou a seleção de $must-link$ e $cannot-link$ na segunda parte da metodologia. Futuramente, POST ajuda a selecionar verbos para caracterizar RS. Igualmente, apoiará a resolução de anáforas, conjugações, sinónimos e inversões presentes nos contextos. Ainda servirá para lidar com relações complexas e relações *multi-hop*.

Entretanto, o tempo gasto nas diferentes combinações de *clustering* e o espaço em memória que este processo ocupa, provocou atrasos na obtenção de resultados. De certa forma, influenciou a possibilidade de serem feitos mais testes de avaliação. Nomeadamente, seriam testes prosseguidos com todos os pares, cujos elementos são EN consecutivas nas respetivas frases, ou com todas as pesagens na 2ª execução. A procura automática do número ótimo de *clusters* a obter reduziria a dimensão desses testes. De qualquer modo, contribuições de computação paralela serão então prestáveis para experiências mais exigentes espacial e temporalmente.

Para processamento de maior volume de dados do que os usados no trabalho de dissertação, o desenvolvimento dum algoritmo de aprendizagem pode ser o mais indicado. Neste caso, uma opção seria a aplicação de aprendizagem supervisionada. Com base num conjunto de documentos anotados, o algoritmo identificaria e classificaria RS. O auxilio de k-Means incremental seria importante.

Apêndice A

Funcionamento do reconhecedor de EN

O reconhecimento das EN é conseguido através da definição de apenas dois argumentos[1]. Estes são *Data* (uma *string* de caracteres) e *type* (um algarismo). Há três combinações válidas de valores para tais argumentos:

- Se $type = 0$, então *Data* é um caminho para uma pasta onde estejam ficheiros que venham a ser processados.
- Se $type = 1$, então *Data* é um caminho para um ficheiro que deverá ser processado.
- Se $type = 2$, então *Data* é o conteúdo dum texto.

O retorno é um quadro de dados com as EN, respetivas desambiguações e posições por parágrafo, frase e ficheiro (a apresentação desta coluna apenas se $type \neq 2$).

Apêndice B

Legenda sobre medidas de distância

Aqui é apresentada uma ampliação da legenda, usada para demonstração dos resultados de avaliação ao *clustering*, alusiva às diferentes medidas de distância. Concretamente, a legenda encontra-se nos gráficos relativos à 2ª execução da metodologia de avaliação. A figura B.1 apresenta essa legenda com mais detalhe.

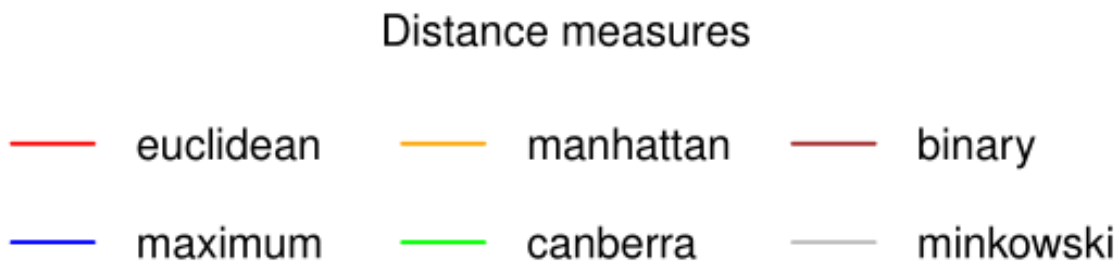


Figura B.1: Legenda que distingue medidas de distância nos gráficos sobre avaliação à aglomeração de RS.

Referências

[1] C. Rocha, *PAMPO: PAMPO - Extract Named Entities from texts*, 2015, r package version 1.15.11.

[2] M. El Enas, "A framework for extracting biological relations from different resources," *International Journal of Computer Applications*, vol. 119, no. 3, 2015.

[3] T. Hasegawa, S. Sekine, and R. Grishman, "Discovering relations among named entities from large corpora," in *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ser. ACL '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. [Online]. Available: http://dx.doi.org/10.3115/1218955.1219008

[4] I. Celik, F. Abel, and G.-J. Houben, "Learning semantic relationships between entities in twitter," in *International Conference on Web Engineering*. Springer, 2011, pp. 167–181.

[5] D. Thenmozhi and C. Aravindan, "An automatic and clause-based approach to learn relations for ontologies," *The Computer Journal*, p. bxv071, 2015.

[6] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the future*, vol. 2007, pp. 1–16, 2012.

[7] P.-N. Tan *et al.*, *Introduction to data mining*. Pearson Education India, 2006.

[8] N. Chinchor and P. Robinson, "Muc-7 named entity task definition," in *Proceedings of the 7th Conference on Message Understanding*, 1997, p. 29.

[9] R. Chaffin, "The concept of a semantic relation," *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pp. 253–288, 1992.

[10] M. Widenius and D. Axmark, *Mysql Reference Manual*, 1st ed., P. DuBois, Ed. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2002.

[11] J. Allen, "Natural language understanding," 1987.

[12] "Yatsko's Computational Linguistics Laboratory," visitado em 21-09-2016. [Online]. Available: http://yatsko.zohosites.com/cll-tagger.html

[13] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4, pp. 175–308, 2006.

[14] B. Percha and R. B. Altman, "Learning the structure of biomedical relationships from unstructured text," *PLoS Comput Biol*, vol. 11, no. 7, p. e1004216, 2015.

[15] L. Dong, F. Wei, H. Sun, M. Zhou, and K. Xu, "A hybrid neural model for type classification of entity mentions," in *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 1243–1249.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[17] K. Lambrou-Latreille, "Relation extraction pattern ranking using word similarity," in *NAACL-HLT 2015 Student Research Workshop (SRW)*, 2015, p. 25.

[18] P. D. Turney, P. Pantel *et al.*, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, no. 1, pp. 141–188, 2010.

[19] F. Petroni, L. Del Corro, and R. Gemulla, "Core: Context-aware open relation extraction with factorization machines," in *Proceedings of EMNLP*, 2015, pp. 1763–1773.

[20] Y. Z. A. JATOWT and K. TANAKA, "Finding"similar"concepts with evidences across different feature vector spaces."

[21] L. Specia, "Translating from complex to simplified sentences," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2010, pp. 30–39.

[22] C. Gasperin, E. Maziero, and S. M. Aluisio, "Challenging choices for text simplification," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2010, pp. 40–50.

[23] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.

[24] J. Silva, A. Branco, S. Castro, and R. Reis, "Out-of-the-box robust parsing of portuguese," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2010, pp. 75–85.

[25] F. de Sa Mesquita, "Extracting Information Networks from Text."

[26] C. Wang, Y. Song, D. Roth, C. Wang, J. Han, H. Ji, and M. Zhang, "Constrained information-theoretic tripartite graph clustering to identify semantically similar relations," in *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 3882–3889.

[27] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering." in *SDM*, vol. 4. SIAM, 2004, pp. 333–344.

[28] I. Feinerer and K. Hornik, *tm: Text Mining Package*, 2015, r package version 0.6-2. [Online]. Available: http://CRAN.R-project.org/package=tm

[29] "Sapo labs." [Online]. Available: http://labs.sapo.pt/

[30] N. Morais, C. Rocha, and A. Jorge, "Framework for clustering of semantic relations," https://github.com/LIAAD/EntityRelationsNelson, 2016.